Current practices in data sharing and challenges ahead!

Timon Oefelein
Senior Manager Account Development,
Strategic Partnerships and Outreach

Tromsø, November 2018

SPRINGER NATURE

The importance of verification

"The strongest arguments provide nothing so long as conclusions are not verified by experience." ***Opus Tertium*, c. 1267**

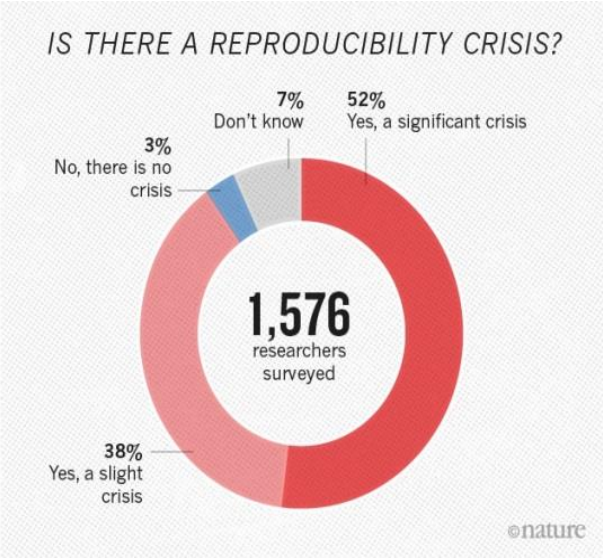
Roger Bacon, 1214-1291

The problem!

Scale and cost of issue

- Pharma reports **75%+ failure rates** [2,3]
- **Irreproducible biology research costs US $28b per year**[1]

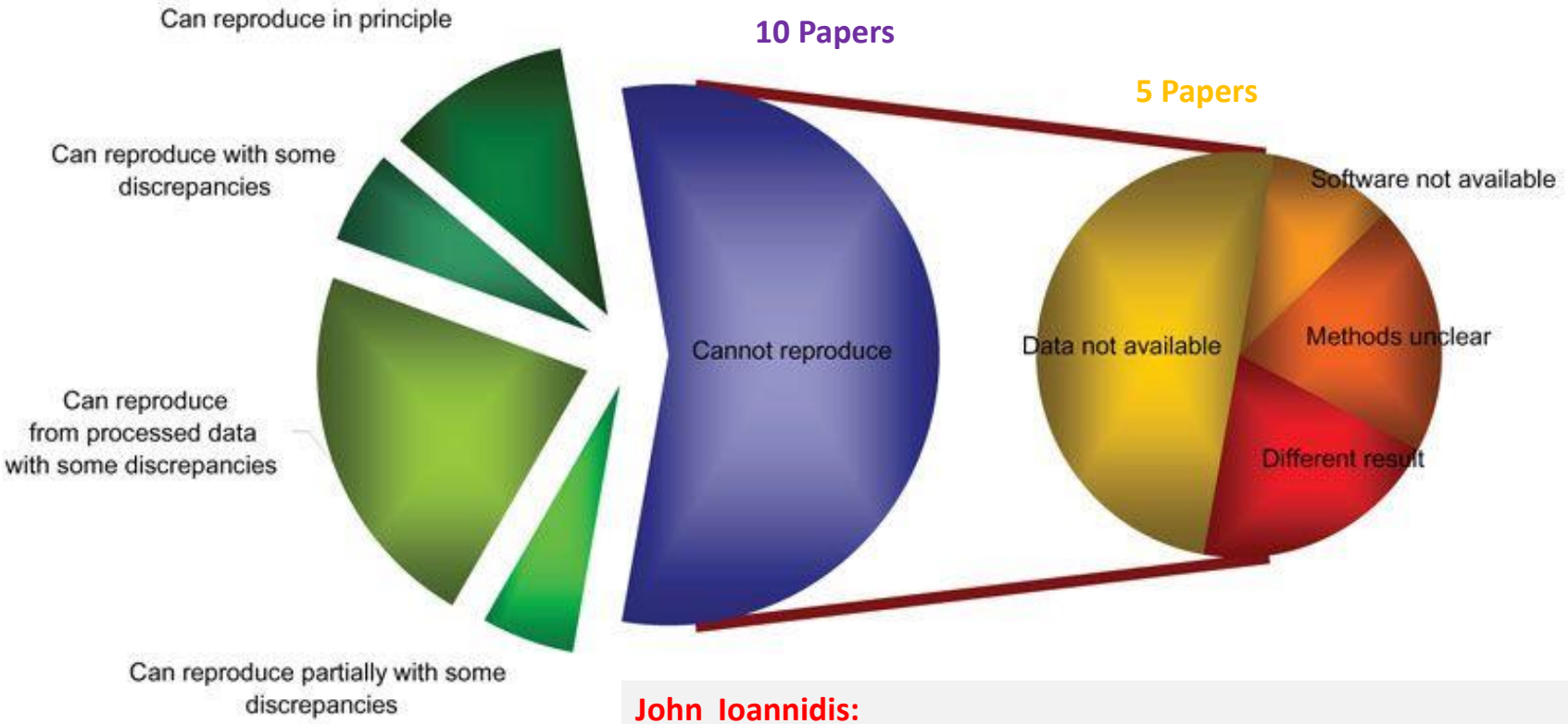
Community perspective



- **52% said "Yes, a significant crises"**
- 50% couldn't reproduce own work
- 70% couldn't reproduce work of others

1. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. *PLoS Biol.* **13**, e1002165 (2015) http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165
2. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012), 3. Prinz, F., Schlange, T. & Asadullah, K. *Nature Rev. Drug Discov.* **10**, 712 (2011)
4. Baker (2015) http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
5. Ioannidis et al (2009) https://www.nature.com/ng/journal/v41/n2/full/ng.295.html

One of the main causes: missing data!



John Ioannidis:

"The main reason for failure to reproduce was data unavailability, and … incomplete data annotation or specification of data processing and analysis."

Nature Genetics Volume 41, ages 149-155, Repeatability of published microarray gene expression analyses. Copyright 2009.

Additional benefits of sharing data =

x2 the publication output of a study

Study: 7000 studies in social & behavioral science, funded by NSF and NIH:

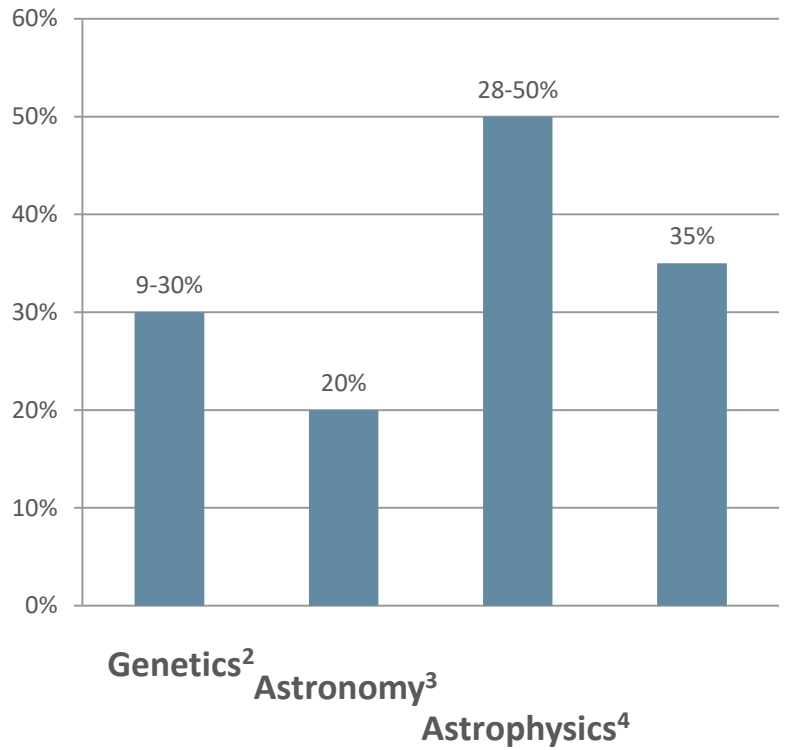
Papers with data:

- **Repository/archived** = **10** (median) publications*
- **Not shared** = **4** (median) publications*[1]

* Resulting publications by PI or a member of Research Team. Excludes secondary publications by non-team members.

up to +50% more citations

Meta-analysis: Articles with data available are cited 9-50% more, depending on the field:



1. Pienta et al (2010) https://deepblue.lib.umich.edu/handle/2027.42/78307
2. Piwowar & Vision (2013) https://doi.org/10.7717/peerj.175
3. Henneken & Accomazzi (2011) https://arxiv.org/abs/1111.3618
4. Dorch et al (2015) https://arxiv.org/abs/1511.02512
5. Sears et al (2011) https://figshare.com/articles/Data_Sharing_Effect_on_Article_Citation_Rate_in_Paleoceanography/1222998/1

Current data sharing practices amongst researchers

2.0

Survey on data sharing practices *(at point of publication)*

- Focus: data sharing at "point of publication"
- Defined sharing: long term (repository or ESM)
- **We covered following aspects:**
- Methodology: survey based
- Participants: 7700 responses, global reach
- White Paper

Importance of data discovery
Portion of researchers actively sharing data
Method of sharing
Size of datasets shared
Obstacles to sharing
Subject & Regional difference to above

Figure 18: Number of respondents by region of the world

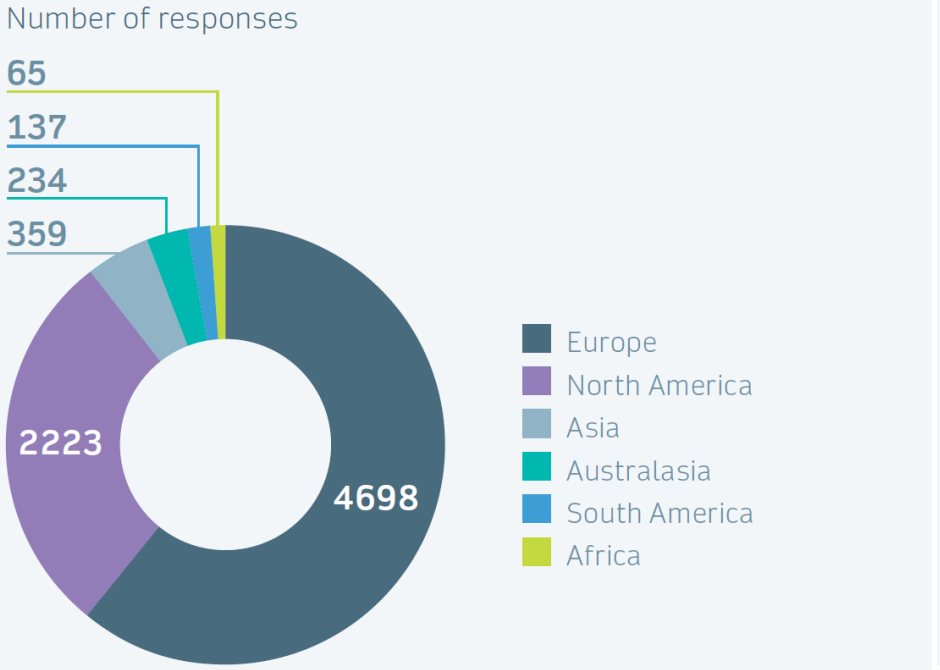
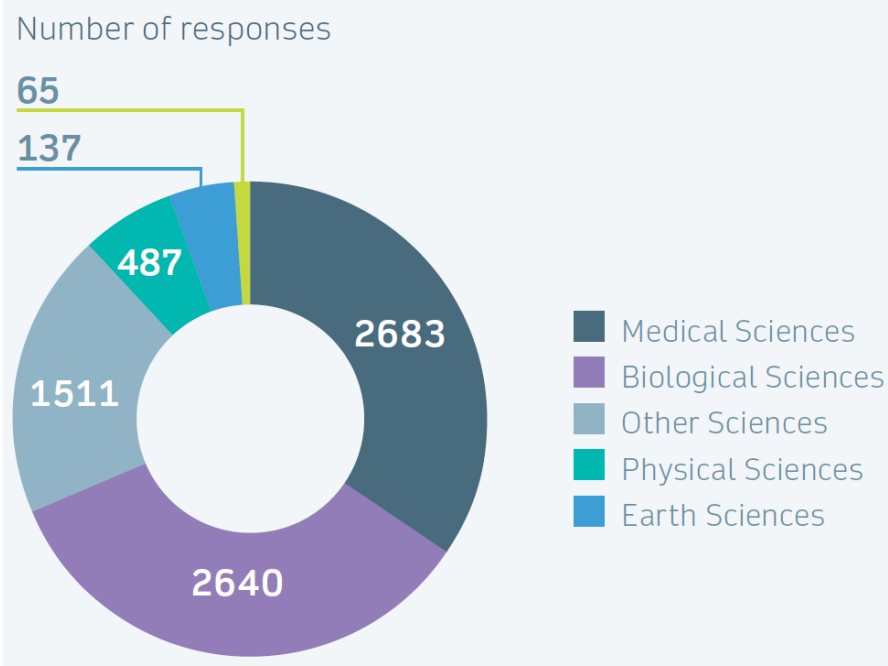
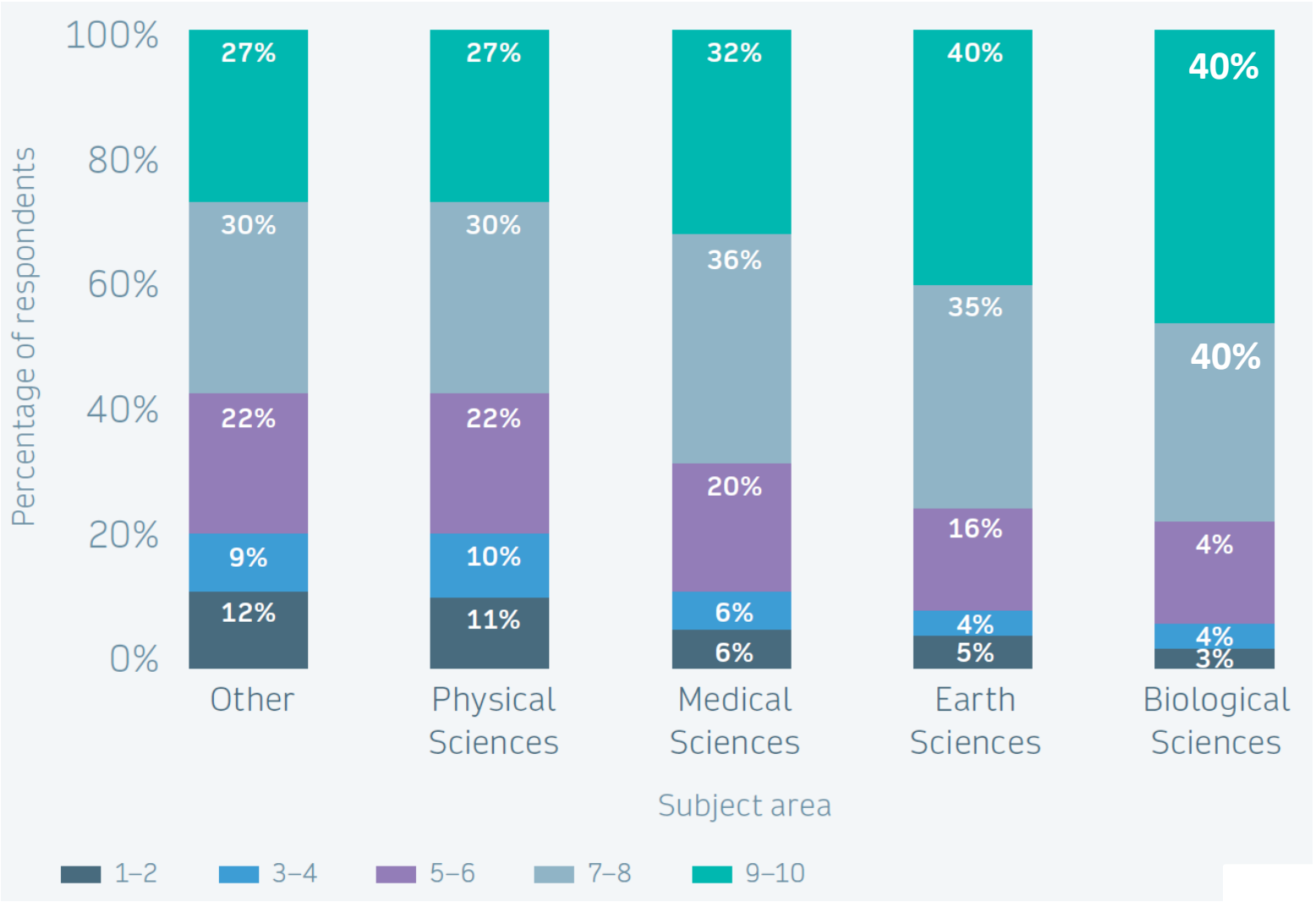


Figure 19: Number of respondents by subject

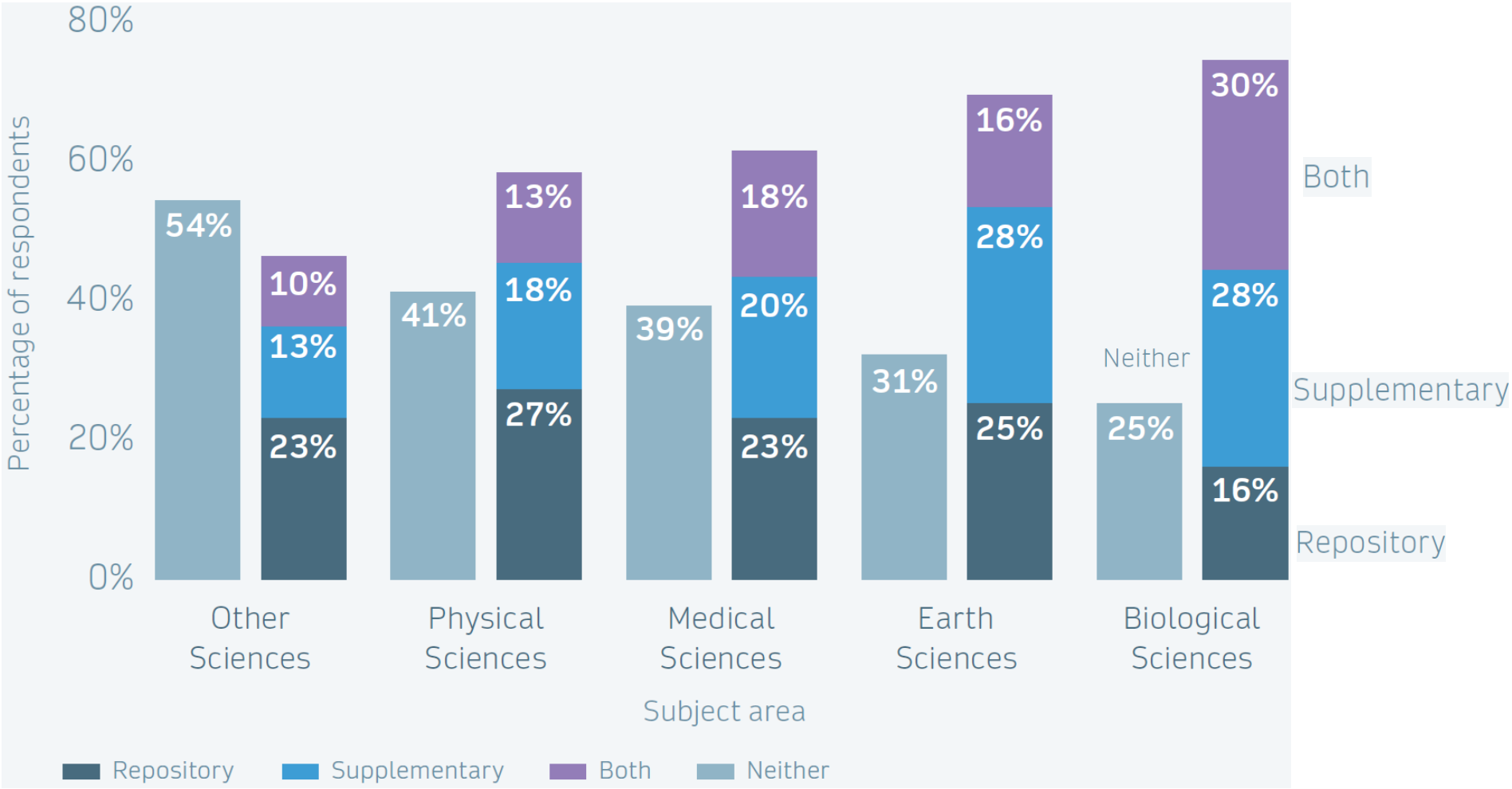


Baynes, Grace, et al. (2018): Whitepaper: Practical challenges for researchers in data sharing. figshare. Paper.

The importance that data are discoverable in different subject areas (1 is the least important, n=7626)



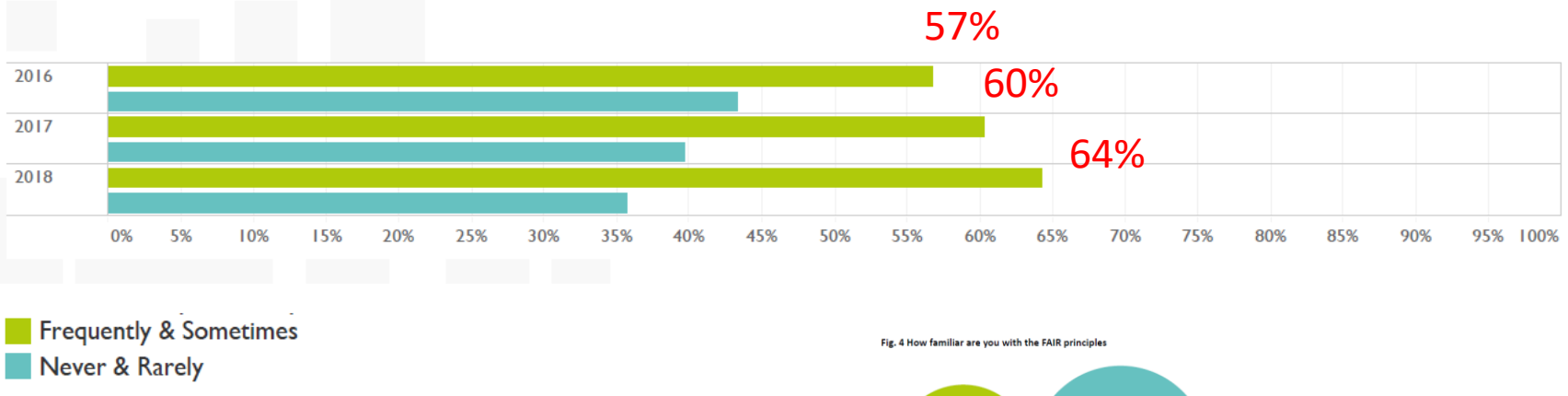
Depositing of data in subject areas (n=7664) *(at point of publication..)*



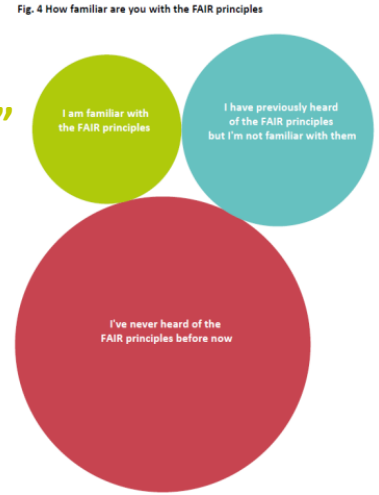
Baynes, Grace, et al. (2018): Whitepaper: Practical challenges for researchers in data sharing. figshare. Paper.

DS: State of open data in 2018 *(before/after publication)*

Fig 1. How often researchers have made their data openly available



"Familiar with FAIR"

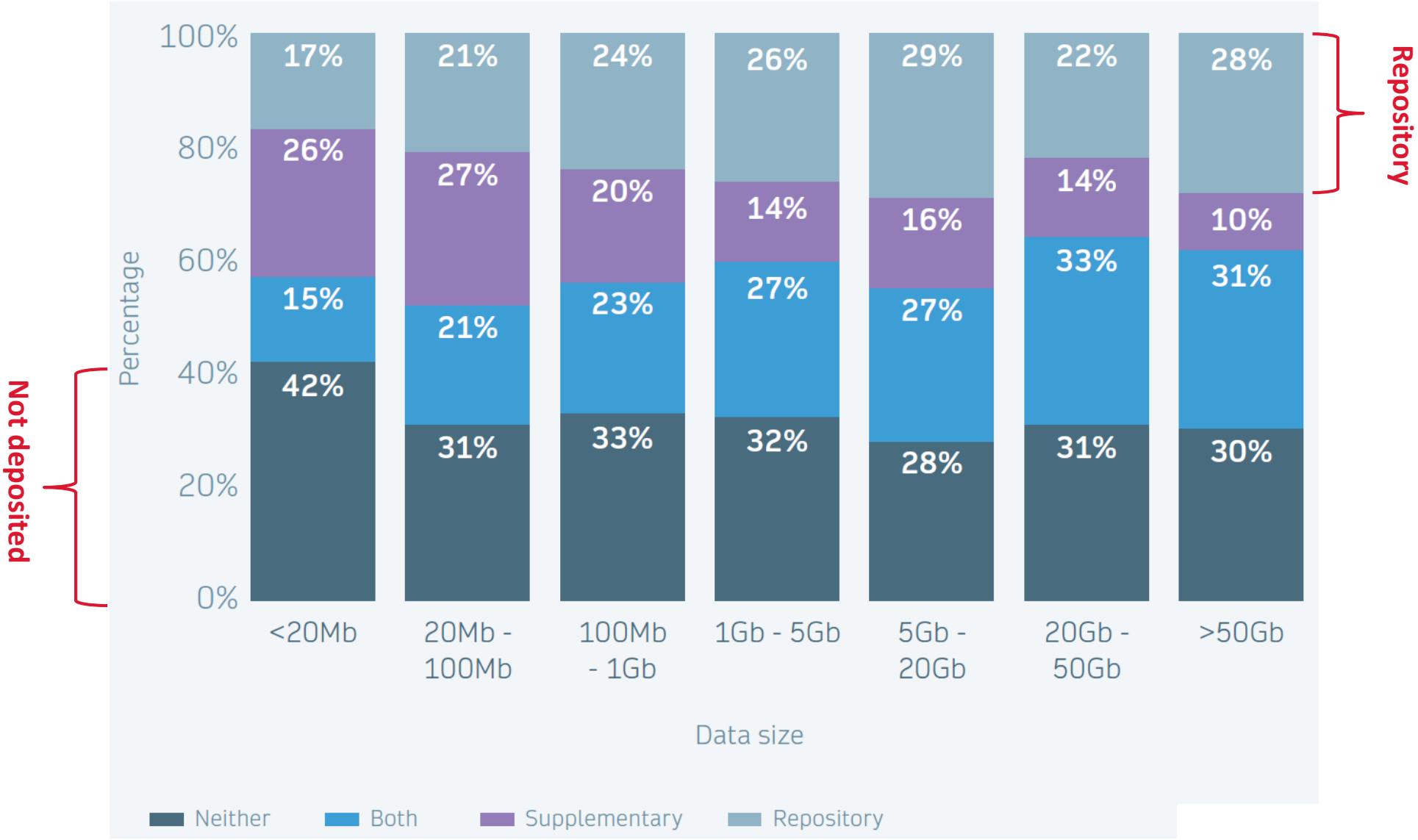


"Heard but not familiar"

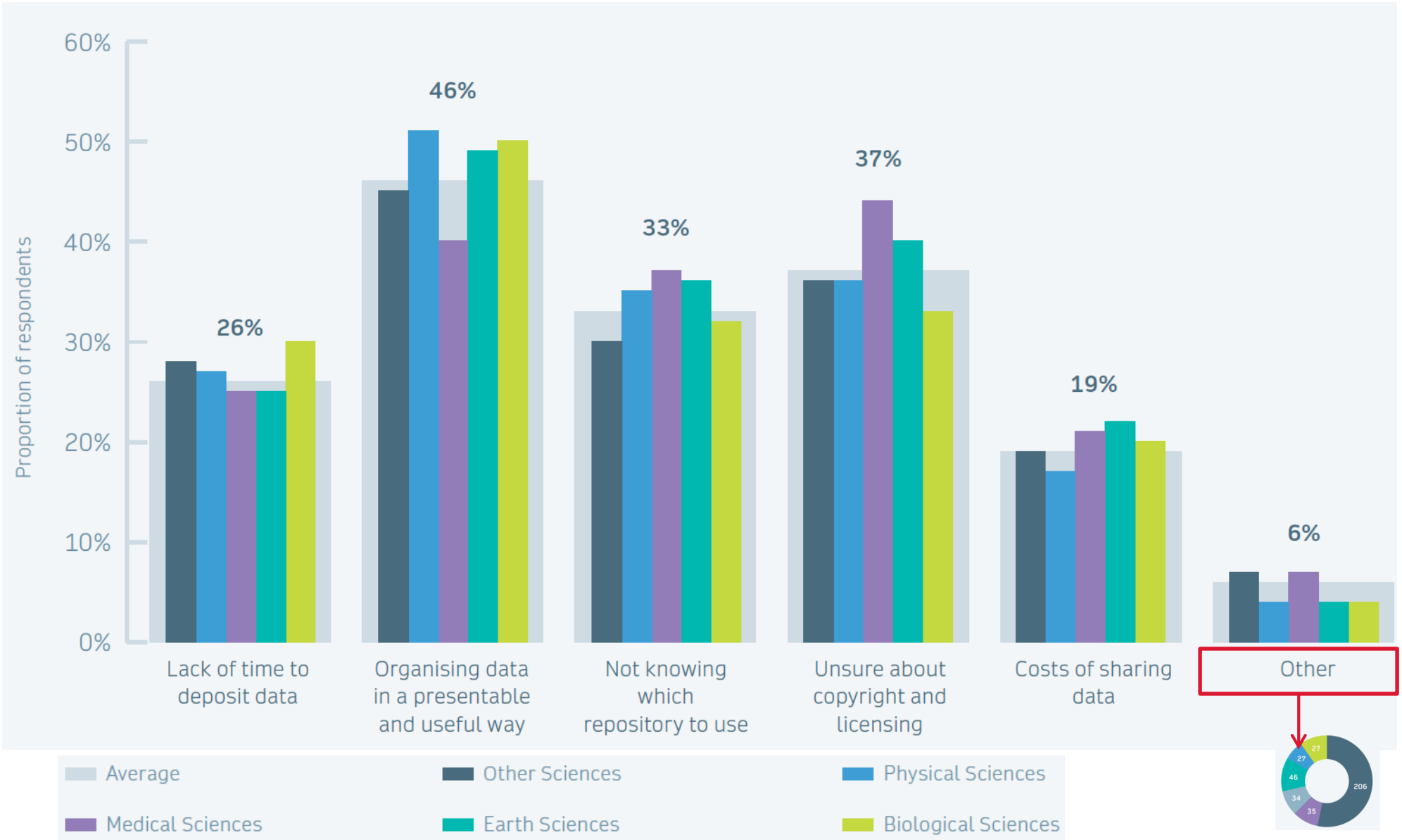
"I've never heard of the FAIR principles before now"

Science, Digital; Hahnel, et al. (2018): The State of Open Data Report 2018. figshare. Paper

Data sharing behaviour by size of dataset (n=6513)



What problems do authors have in sharing datasets?



What can publishers do to help researchers overcome these problems?

3.0

How are publishers responding to these challenges?

SPRINGER NATURE:

- Research Data Help Desk
- Recommended repositories
- Data policies
- Helping set standards:
- Implement citation and linking practices
- Research Data Service
- New: Data Availability Reporting
- Credit for data sharing via badges

SCIENTIFIC DATA

RDA
RESEARCH DATA ALLIANCE

DataCite

Springer Nature research data policy initiative

Type 1

Encourages:
- Data sharing
- Data citation

Type 2

Encourages:
- Data sharing
- Data citation
- Data-Availability-Statement

Type 3

Assumes:
- Data-sharing

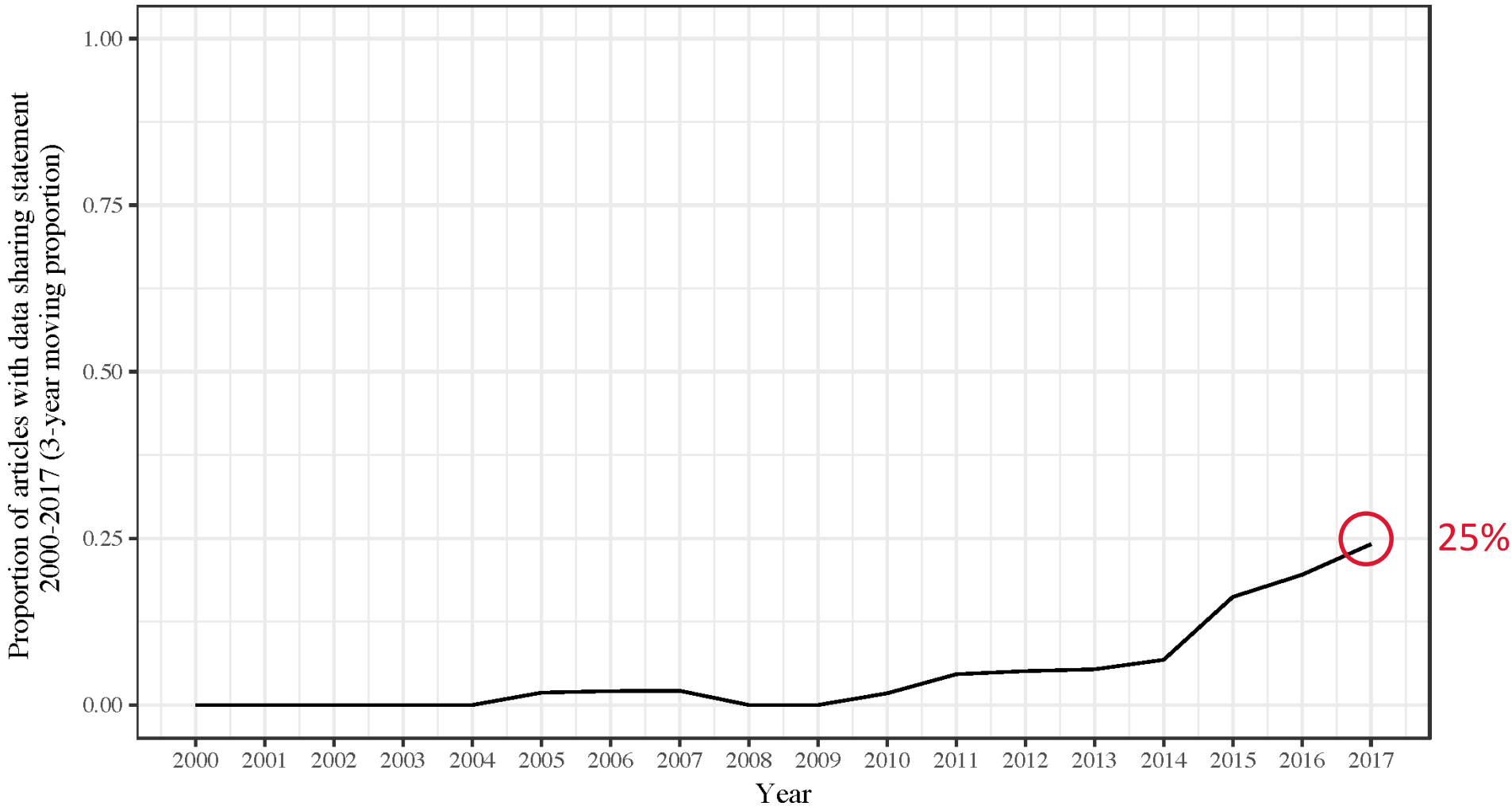
Mandatory:
- Data Avail-Statement
- Data citation

Type 4

Mandatory:
- Data sharing (reader + reviewer)
- Data-Avail-Statement
- Data citation

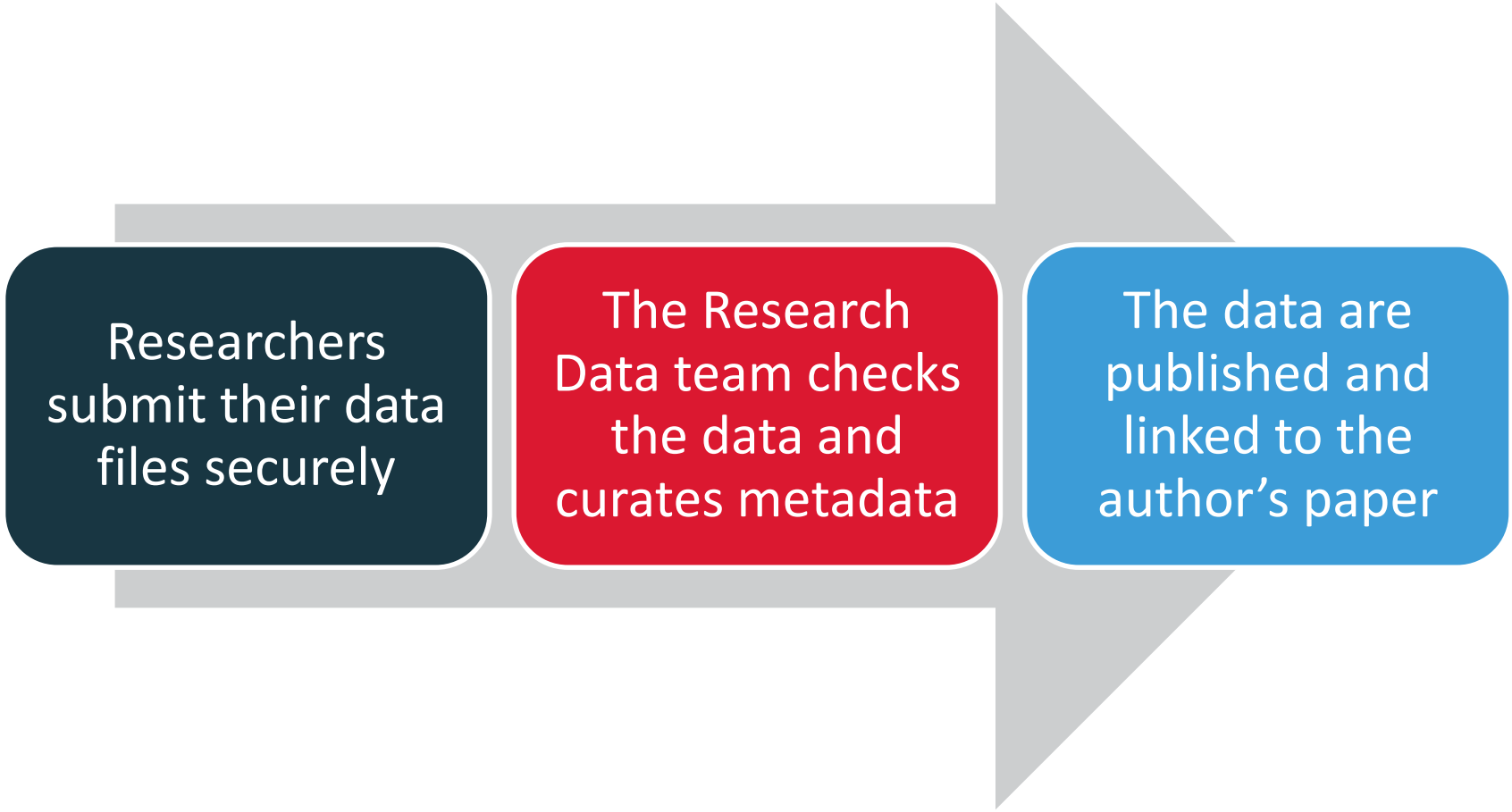
- SRG journals: **over 1,400** (>50%) adopted a policy
- Nature journals: most are Type 3
- Preference for **repositories over Electronic Supplementary Material**

Proportion of articles with data sharing statement



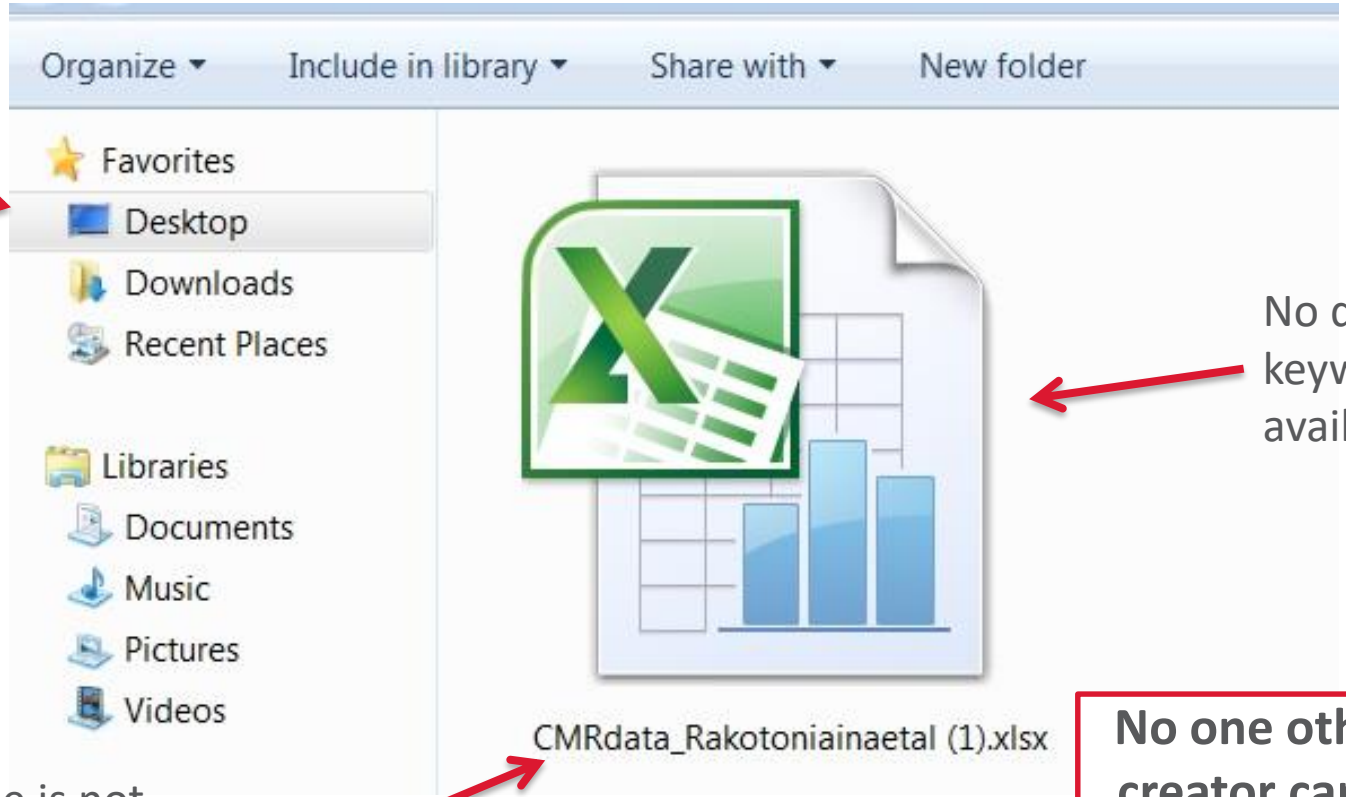
J.D. Wallach et al., "Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017," *PLOS Biol*, 16:e2006930, 2018.

Helpdesk and Data Curation Service



Before data curation: a researcher's dataset in a desktop folder

The dataset is stored as an Excel file in a desktop folder



No description or keywords available

The file title is not comprehensible to anyone but the creator

No one other than the creator can access the data, or even knows that it exists

Before curation begins

Once received, we check to make sure that the dataset is suitable for our curation services. Multiple files in any format are accepted.

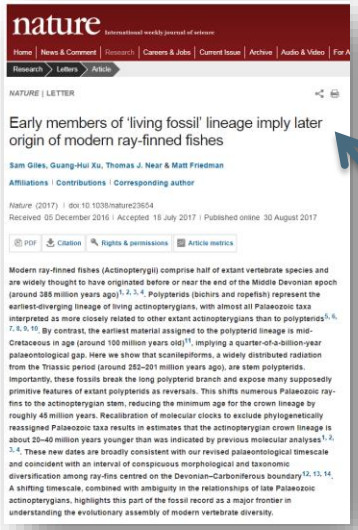
Pre-curation data checks:

- ✓ The data aren't sensitive
- ✓ The data don't include direct or indirect human identifiers
- ✓ The data shouldn't be in a community repository
- ✓ The data are associated with a trusted publication

After making these checks, we begin the curation process. If necessary we may recommend that the dataset is split into smaller groups or collections.

Example of output of Research Data Support

Paper published in **Nature**



Early members of 'living fossil' lineage imply later origin of modern ray-finned fishes

Sam Giles, Guang-Hui Xu, Thomas J. Near & Matt Friedman

Affiliations | Contributions | Corresponding author

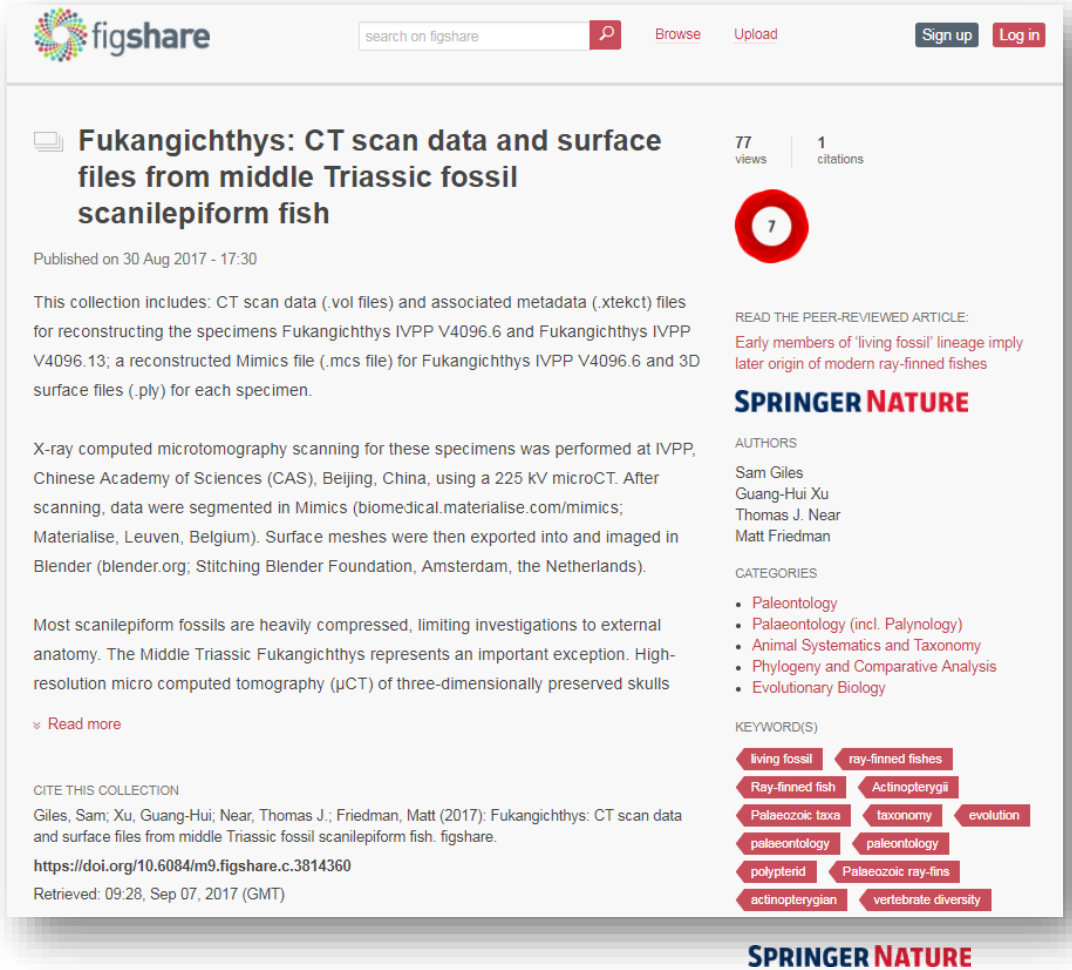
Nature (2017) | doi:10.1038/nature23654

Received 05 December 2016 | Accepted 18 July 2017 | Published online 30 August 2017

Modern ray-finned fishes (Actinopterygii) comprise half of extant vertebrate species and are widely thought to have originated before or near the end of the Middle Devonian epoch (around 385 million years ago)[1, 2, 3, 4]. Polypterids (bichirs and ropefish) represent the earliest-diverging lineage of living actinopterygians, with almost all Palaeozoic taxa interpreted as more closely related to other extant actinopterygians than to polypterids[5, 6, 7, 8, 9, 10]. By contrast, the earliest material assigned to the polypterid lineage is mid-Cretaceous in age (around 100 million years old)[11], implying a quarter-of-a-billion-year palaeontological gap. Here we show that scanilepiforms, a widely distributed radiation from the Triassic period (around 252–201 million years ago), are stem polypterids. Importantly, these fossils break the long polypterid branch and expose many supposedly primitive features of extant polypterids as reversals. This shifts numerous Palaeozoic ray-fins to the actinopterygian stem, reducing the minimum age for the crown lineage by roughly 45 million years. Recalibration of molecular clocks to exclude phylogenetically reassigned Palaeozoic taxa results in estimates that the actinopterygian crown lineage is about 20–40 million years younger than was indicated by previous molecular analyses[1, 2, 3, 4]. These new dates are broadly consistent with our revised palaeontological timescale and coincident with an interval of conspicuous morphological and taxonomic diversification among ray-fins centred on the Devonian–Carboniferous boundary[12, 13, 14]. A shifting timescale, combined with ambiguity in the relationships of late Palaeozoic actinopterygians, highlights this part of the fossil record as a major frontier in understanding the evolutionary assembly of modern vertebrate diversity.

Data availability statement included with the paper

Data published in **Figshare**



figshare

search on figshare

Browse Upload Sign up Log in

Fukangichthys: CT scan data and surface files from middle Triassic fossil scanilepiform fish

77 views | 1 citations

Published on 30 Aug 2017 - 17:30

This collection includes: CT scan data (.vol files) and associated metadata (.xtekct) files for reconstructing the specimens Fukangichthys IVPP V4096.6 and Fukangichthys IVPP V4096.13; a reconstructed Mimics file (.mcs file) for Fukangichthys IVPP V4096.6 and 3D surface files (.ply) for each specimen.

X-ray computed microtomography scanning for these specimens was performed at IVPP, Chinese Academy of Sciences (CAS), Beijing, China, using a 225 kV microCT. After scanning, data were segmented in Mimics (biomedical.materialise.com/mimics; Materialise, Leuven, Belgium). Surface meshes were then exported into and imaged in Blender (blender.org; Stitching Blender Foundation, Amsterdam, the Netherlands).

Most scanilepiform fossils are heavily compressed, limiting investigations to external anatomy. The Middle Triassic Fukangichthys represents an important exception. High-resolution micro computed tomography (μCT) of three-dimensionally preserved skulls

Read more

CITE THIS COLLECTION

Giles, Sam; Xu, Guang-Hui; Near, Thomas J.; Friedman, Matt (2017): Fukangichthys: CT scan data and surface files from middle Triassic fossil scanilepiform fish. figshare.

https://doi.org/10.6084/m9.figshare.c.3814360

Retrieved: 09:28, Sep 07, 2017 (GMT)

READ THE PEER-REVIEWED ARTICLE:

Early members of 'living fossil' lineage imply later origin of modern ray-finned fishes

SPRINGER NATURE

AUTHORS

Sam Giles
Guang-Hui Xu
Thomas J. Near
Matt Friedman

CATEGORIES

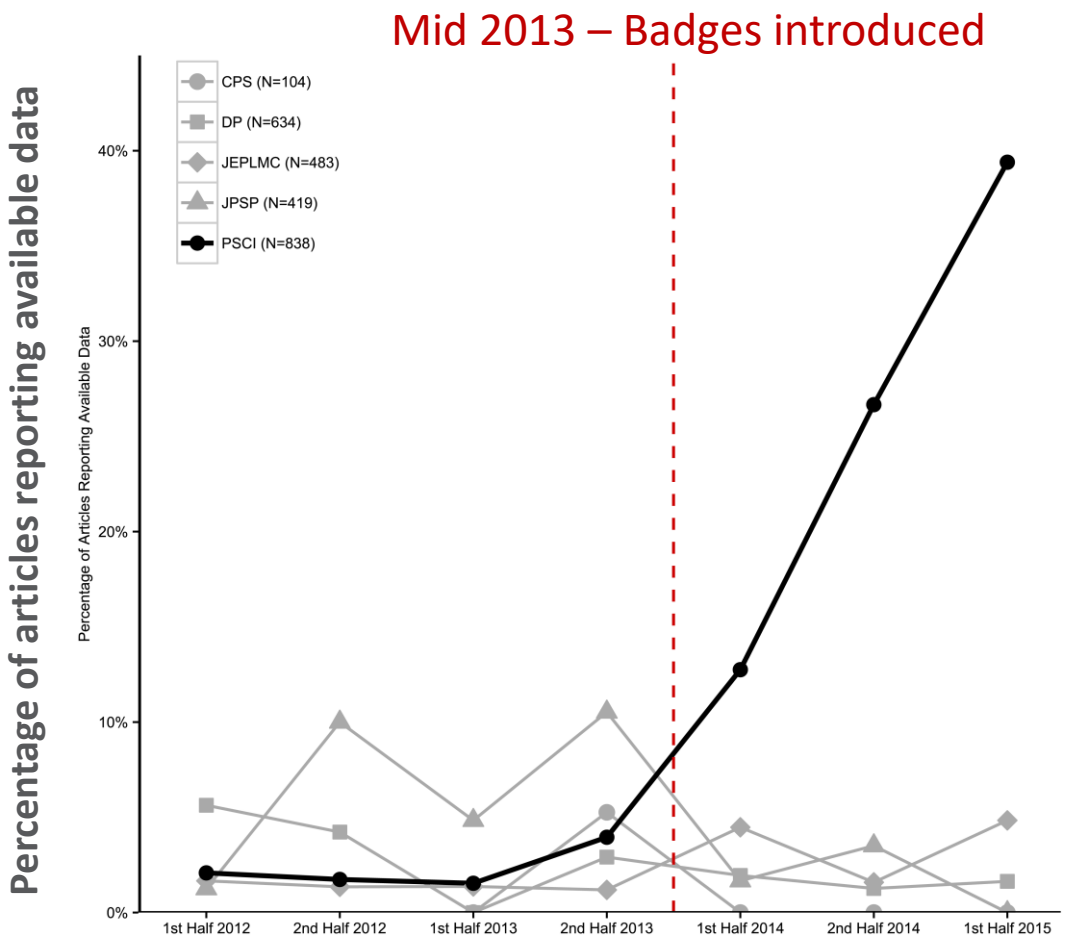
- Paleontology
- Palaeontology (incl. Palynology)
- Animal Systematics and Taxonomy
- Phylogeny and Comparative Analysis
- Evolutionary Biology

KEYWORD(S)

living fossil | ray-finned fishes | Ray-finned fish | Actinopterygii | Palaeozoic taxa | taxonomy | evolution | palaeontology | paleontology | polypterid | Palaeozoic ray-fins | actinopterygian | vertebrate diversity

SPRINGER NATURE

Open Badges Pilot – *BMC Microbiology*



Psychological Science

- **Before badge:** 7% % of art. reported open data
- **After 6 months:** 23% articles “”
- **After 12 months:** 39% articles “”

Badge Criteria (with Center for Open Science):

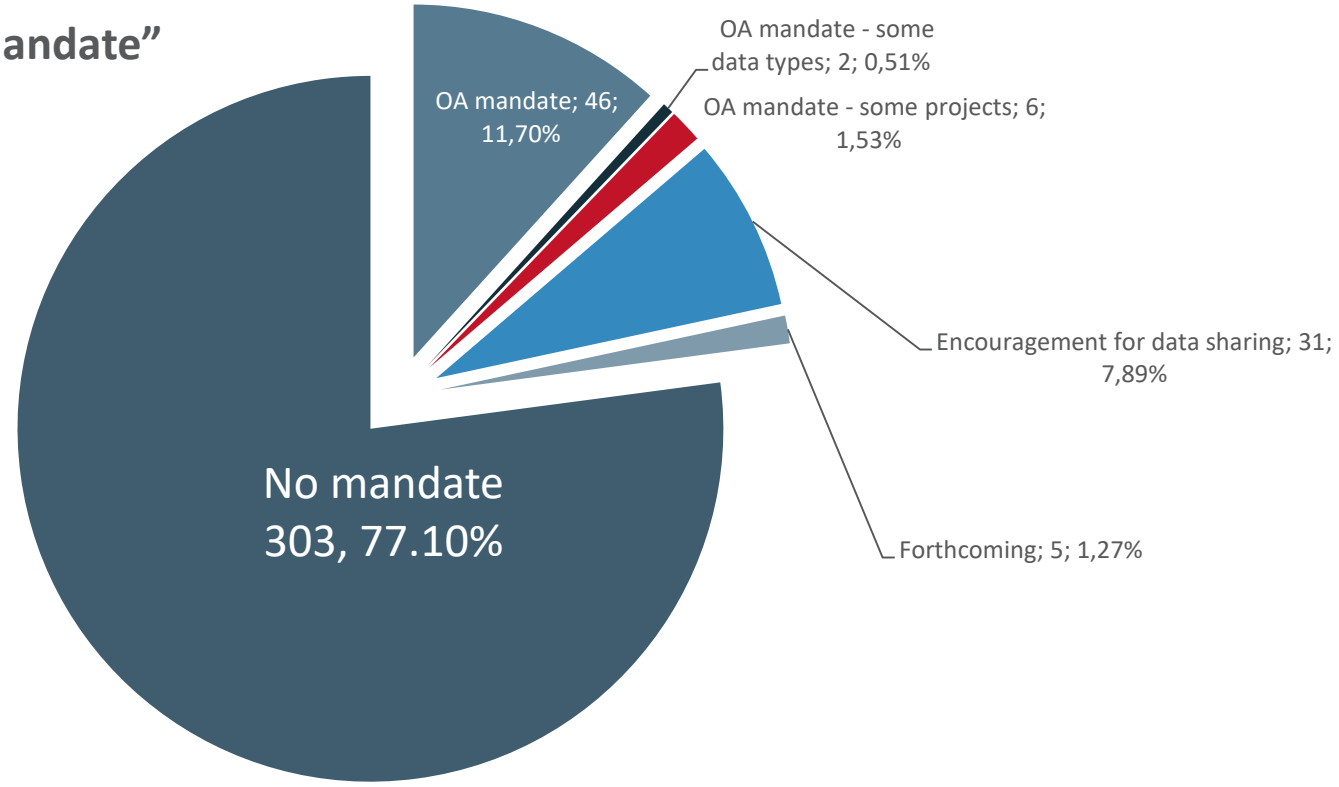
- Data availability statement
- Dataset in a public repository
- Persistent identifier (DOI, Accession number)
- Dataset checked/confirmed as relevant to paper

Kidwell MC, Badges to Acknowledge Open Practices. PLOS Biology 14(5): e1002456. https://doi.org/10.1371/journal.pbio.1002456

Some funders require sharing of research data

393 Funders surveyed (157 from Europe):

- **46 (22%) funders "mandate" data sharing**
- **31 (8%) encourage data sharing**
- **303 (77%) have "no mandate"**



Springer Nature analysis 2018; n = 393

SPRINGER NATURE

Additional key developments:

1. **Researchers citing data** in References/DAS, via DataCite standard
2. **Interconnectivity of repositories**, e.g DANS, dataMED

- **Data search engines,** eg Google Dataset Search
- **Metadata standards,** eg DataCite, Schema.org, Scholix
- **Best practices for data archiving/sharing**, e.g. FAIR
- **Standardization of** data policies, e.g. RDA
- **Growing community** of advocates, e.g. FORCE11
- **Indexes of data repositories**, e.g. re3data.org
- **Data Citation Index** e.g. Clarivate's DCI, Scholix Framework
- **Growth of institutional repositories**

- **EU New data infrastructure pilots..**
  - European Open Science Cloud (€30m)
  - OpenAIRE-Advance (€10m)

SPRINGER NATURE

Thank You!

Nature, Robust research: Institutions must do their part for reproducibility, 1 Sept 2015. Illustration by David Parkins

Thank you

Timon.Oefelein@springernature.com

The story behind the image

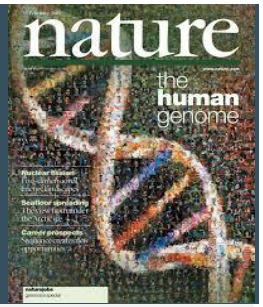
Antarctica meltdown could double sea level rise

Researchers at Pennsylvania State University have been considering how quickly a glacial ice melt in Antarctica would raise sea levels. By updating models with new discoveries and comparing them with past sea-level rise events they predict that a melting Antarctica could raise oceans by more than 3 feet by the end of the century if greenhouse gas emissions continued unabated, roughly doubling previous total sea-level rise estimates. Rising seas could put many of the world's coastlines underwater or at risk of flooding and storm surges.

SPRINGER NATURE

Two success stories!

Human Genome Project



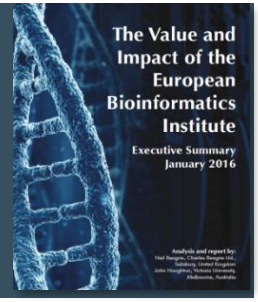
Cost of sequencing human genome fallen from $ 1 Billion in 1990s to $3-500 today.

Nr. of diseases with an identified genomic cause risen from 61 in 1990 to 5000 today.

More than 100 drugs currently on the market have a pharmacogenomics label which facilitates personal medicines.

Data is openly accessible

European Bioinformatics Institute



£47 million: Annual operating cost

£1 billion: Annual efficiency savings to researchers worldwide[2]

£920 million: Estimated annual estimate of future research impacts

Data is openly accessible

1. http://www.unitedformedicalresearch.com/advocacy_reports/the-impact-of-genomics-on-the-u-s-economy
2. http://www.ebi.ac.uk/about/news/press-releases/value-and-impact-of-the-european-bioinformatics-institute

Towards transparency in research reporting

2013: Advocate/Encourage:

- Raise standards of reporting on **statistics:**
- Ensure reporting on: **sample size estimation, randomization, blinding, handling of data**
- Information on **materials** used
- Eliminate length restrictions for **Methods** sections
- Encourage linked **step-by-step protocols** on Protocol Exchange

In 2017: Formal reporting:

- 1. Reporting Summary and 2. Editorial Policy Checklist
- **The Reporting Summary is published with the manuscript**
- Authors fill them out prior to peer review and update them on resubmission
- Available to referees during peer review

SPRINGER NATURE

Example author feedback report

SPRINGER NATURE

Springer Nature Research Data Support feedback

Job number: RDS-XXXX-XXXXX

To improve the discoverability and accessibility of your data, we have made the enhancements your data record in springernature.figshare.com:

- ✓ Copy edited the title
- ✓ Checked/corrected author name and order
- ✓ Copy edited/enhanced your description
- ✓ Added keywords and categories to improve discoverability
- ✓ Added links to your publication
- ✓ Added funding information
- ✓ Unpackaged your data files and uploaded as individually-previewable file

Following deposition and curation, your data now features the following benefits Your data:

- have been assigned a unique identifier by the host repository https://doi.org/10.6084/m9.figshare.**XXXXXX**
- can be previewed in the repository using a web browser
- will be clearly and appropriately linked to the associated publication
- have been assigned a licence that will permit wide reuse
- are easily citable via the figshare repository

Your data record:

- includes a useful, descriptive title
- includes keywords, categories and other metadata that will aid discovery
- is indexed and searchable online
- has a unique, persistent link
- contains descriptive details that can be easily viewed and retrieved within the repository
- describes the data files in sufficient detail to make them easily understood

Areas that could still be improved:

- Your dataset appears to be incomplete based on the description and other data files; it appears that ... are missing. Please check that all relevant data files are present.
- Your data record would benefit from a more detailed description of the underlying methods used to generate these data

Specific guidance from our Research Data Editors:

In the description are you referring to file XXX.xls when you refer to 'data file 1', or to file XXY.xls? I have edited the description to include a more explicit reference to this.

DOI and suggested data citations provided

Useful and meaningful title

Author names consistent

Comprehensive description including the data context of the study and data gathering method

Funder information available

Altmetrics provides information on downloads and citations

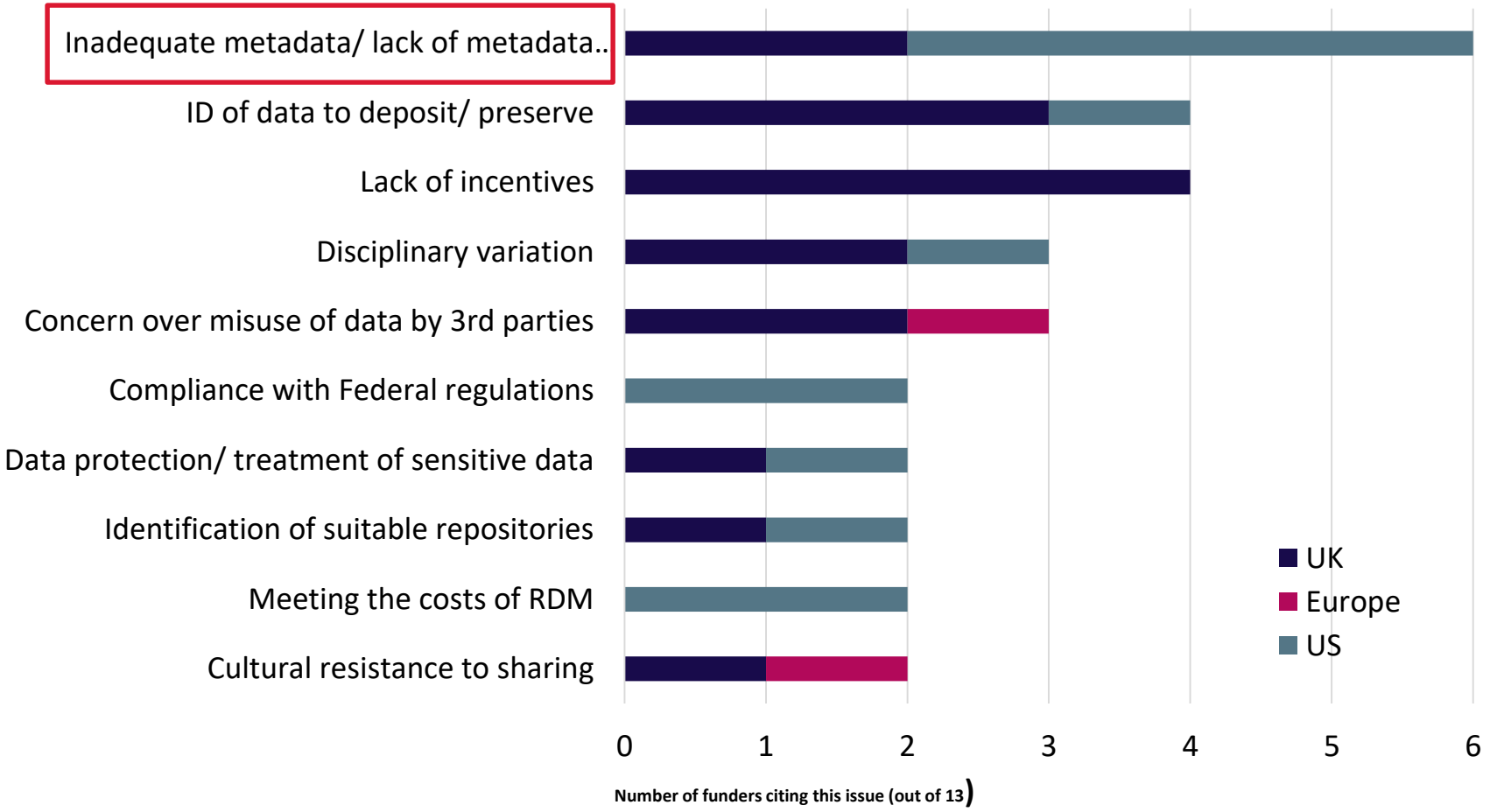
Link to the peer reviewed article connected to this dataset

Relevant categories and keywords allow other researchers to find the dataset

Licence added to make reuse conditions clear

Source: https://doi.org/10.6084/m9.figshare.5259415

Number of funders citing this issue, n= 13



Springer Nature research 2017; n = 13