This slides: https://osf.io/2pdxv

CENTER FOR OPEN SCIENCE

Challenges and opportunities with data sharing

Tim Errington

Center for Open Science

http://cos.io/

JOHN TEMPLETON FOUNDATION

Templeton Religion Trust

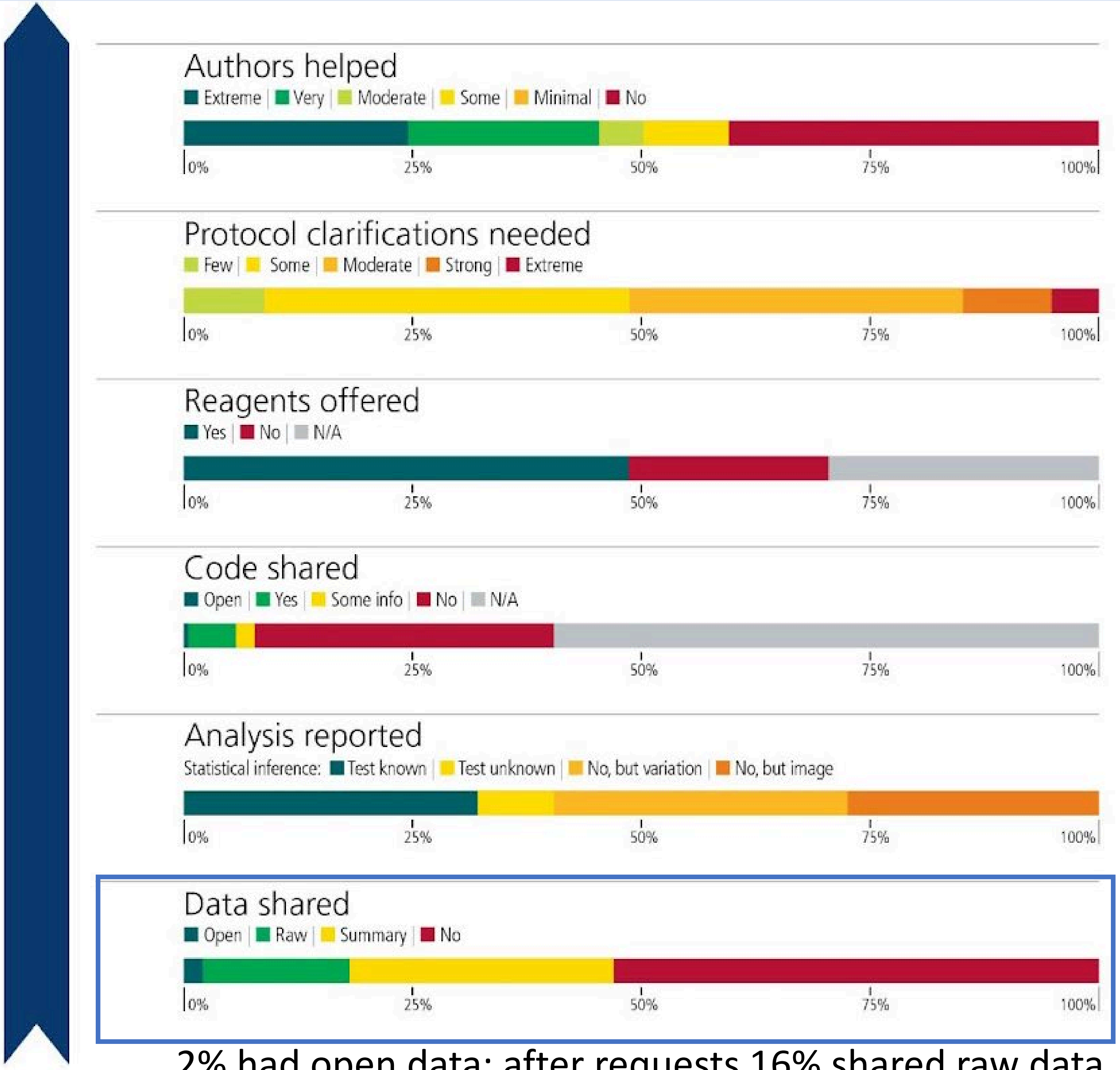
TEMPLETON World Charity Foundation

Outline

- Rates of data sharing
- Attitudes of authors towards data sharing
- Behaviors of data sharing
- FAIR data sharing
- Challenges of data sharing
- Opportunities

How often were data shared?

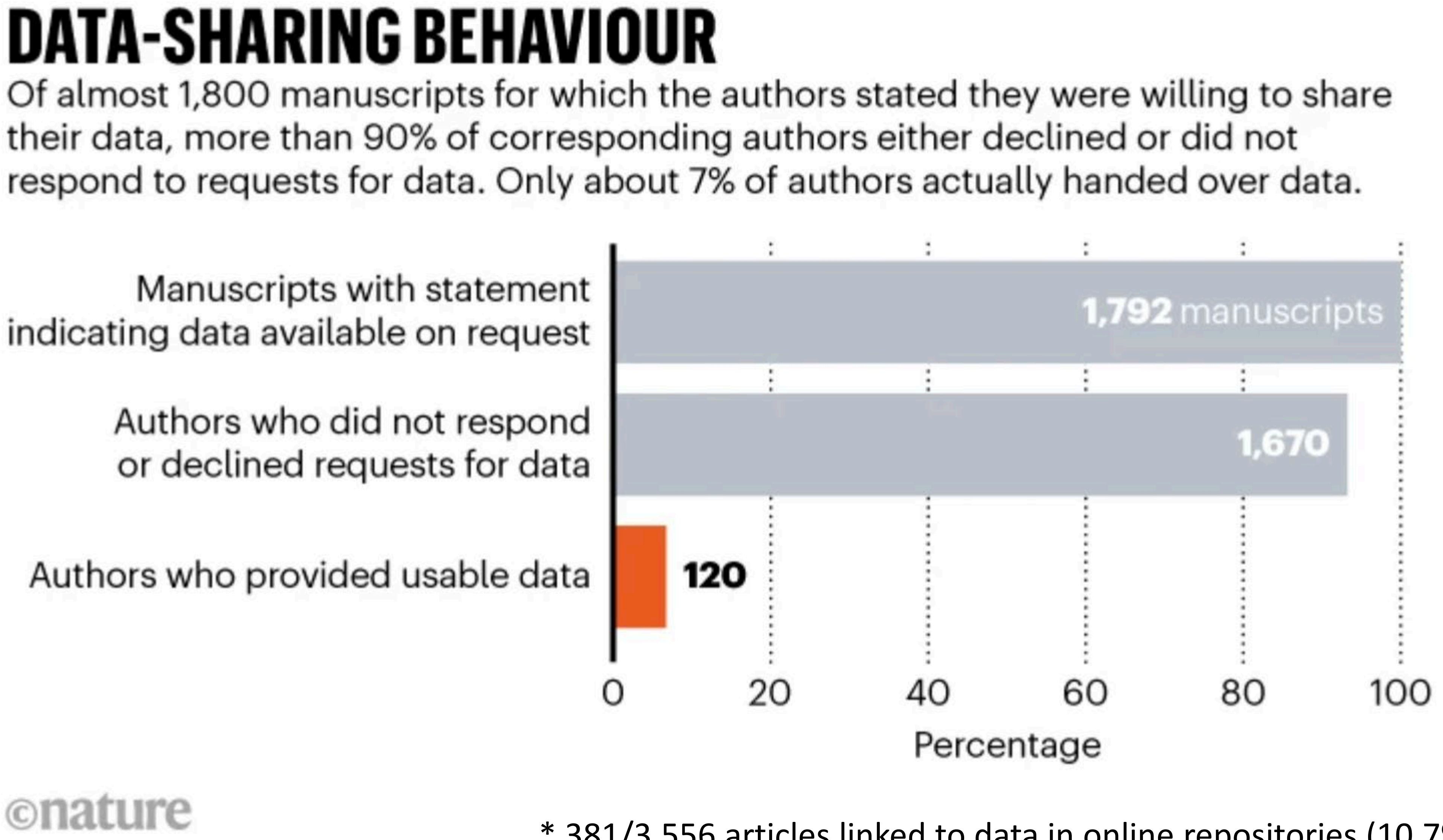
DESIGNED
193 experiments



Errington et al., 2021

2% had open data; after requests 16% shared raw data

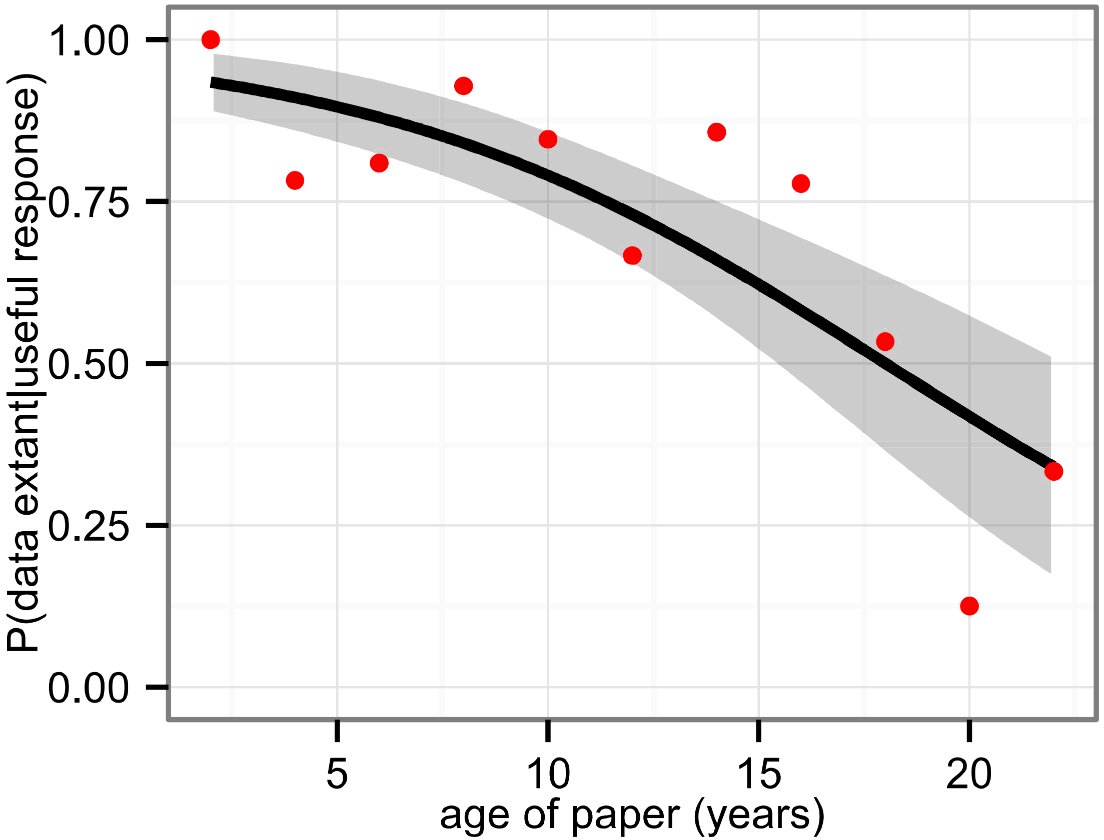
How often were data shared?

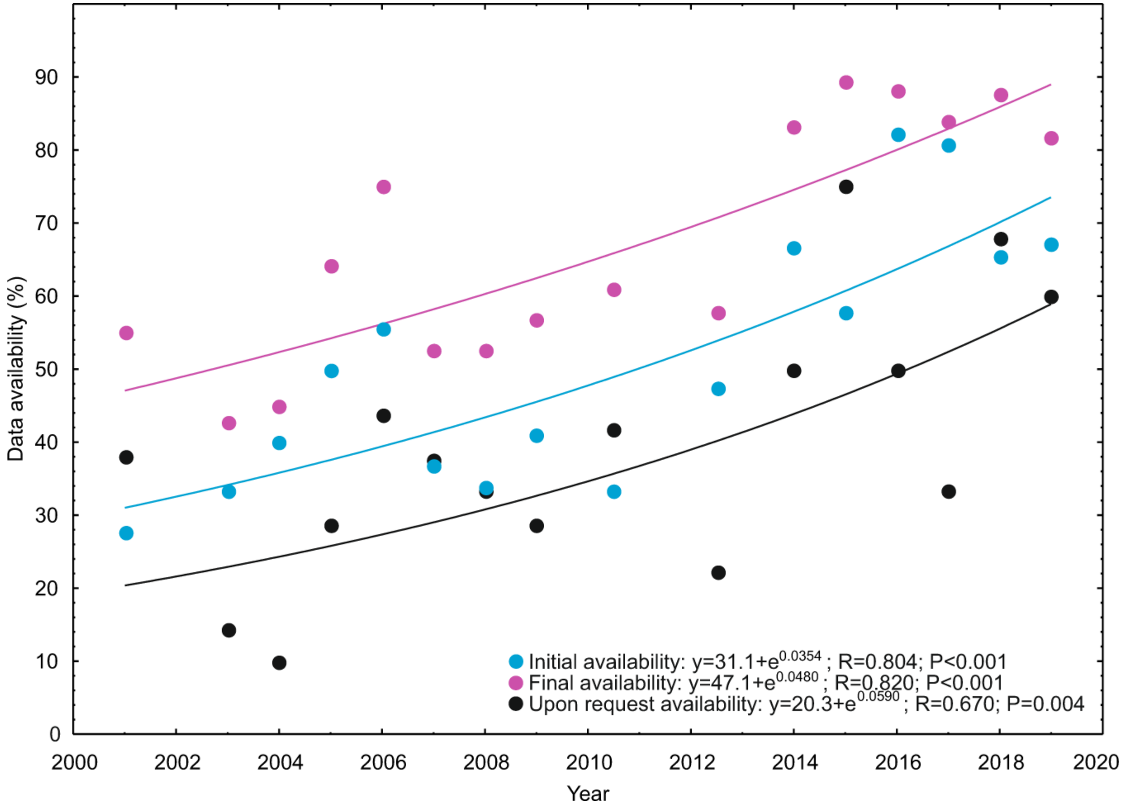


* 381/3,556 articles linked to data in online repositories (10.7%)

Gabelica et al., 2022; Watson, 2022

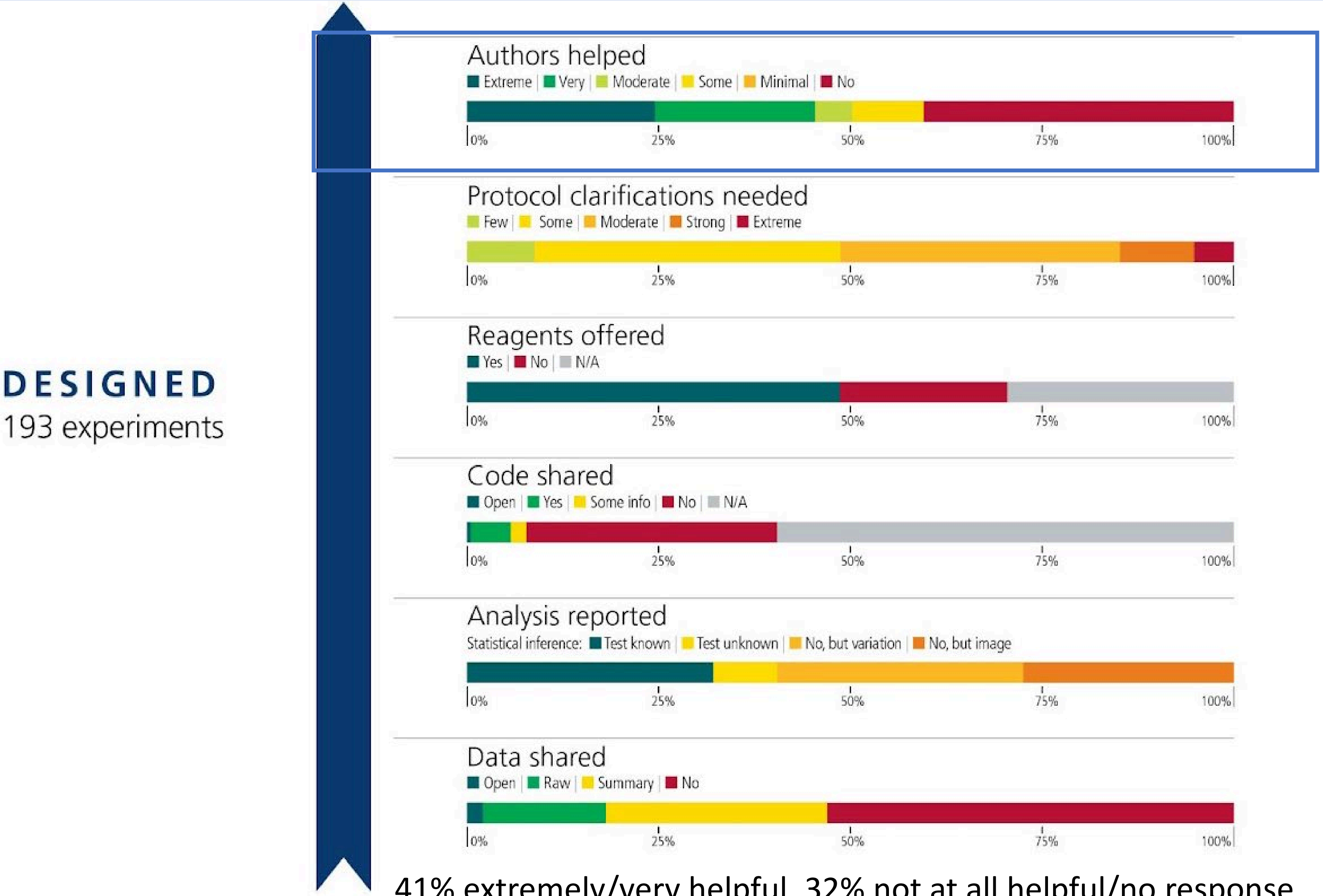
Data access declines with age





Vines et al., 2013; Tedersoo et al., 2021

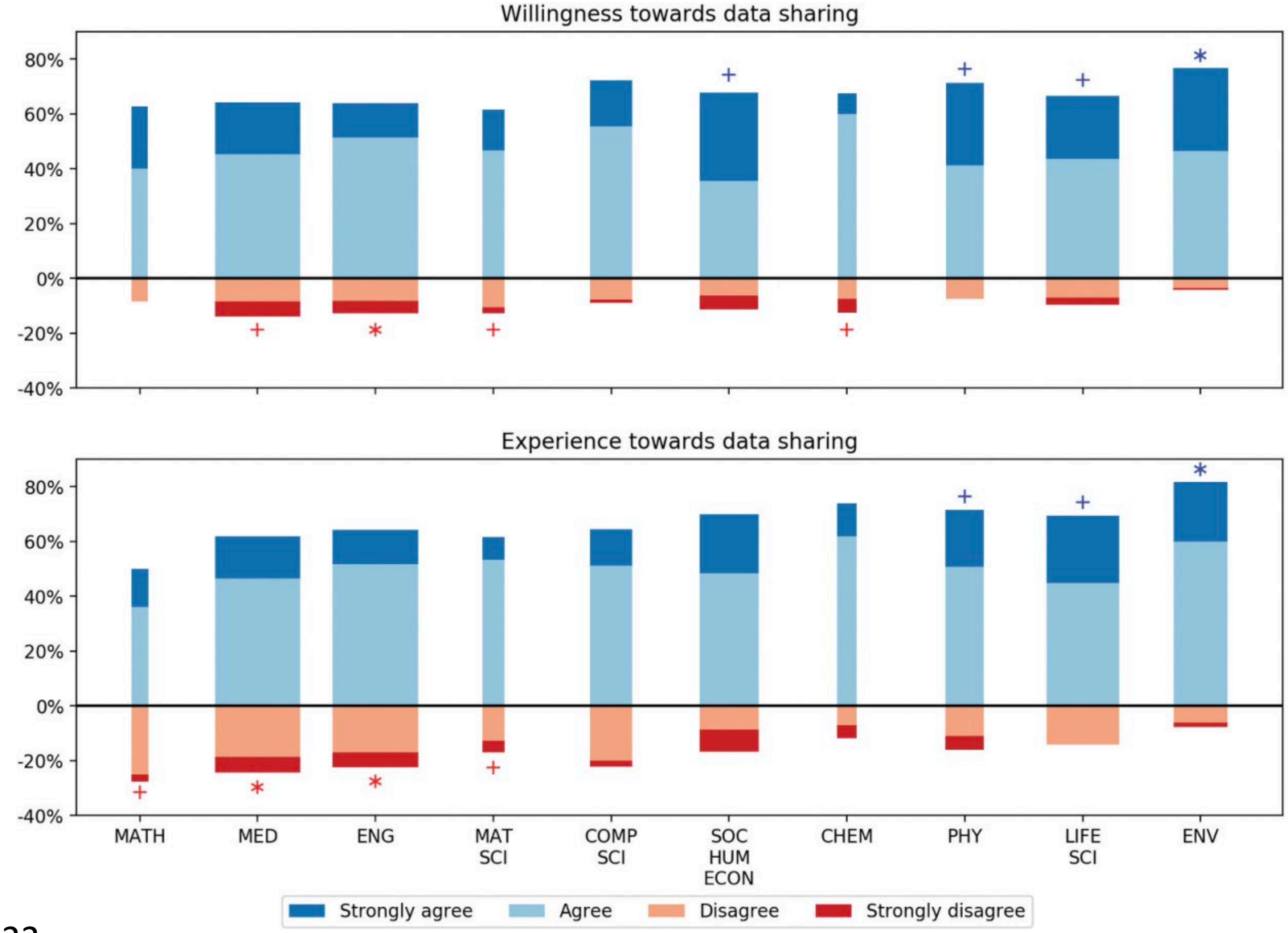
How often was help provided?



41% extremely/very helpful, 32% not at all helpful/no response

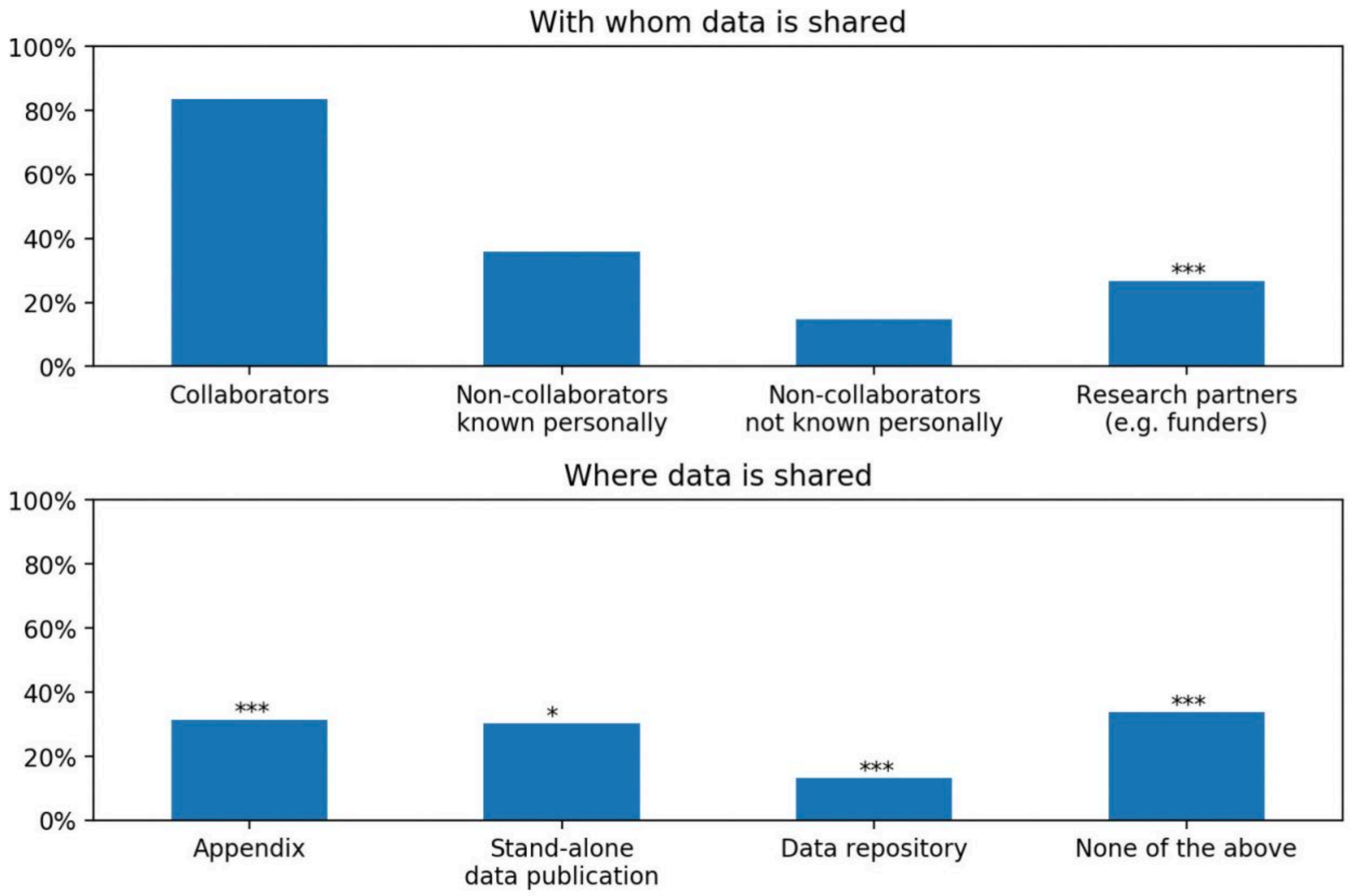
Errington et al., 2021

Attitudes towards data sharing by discipline



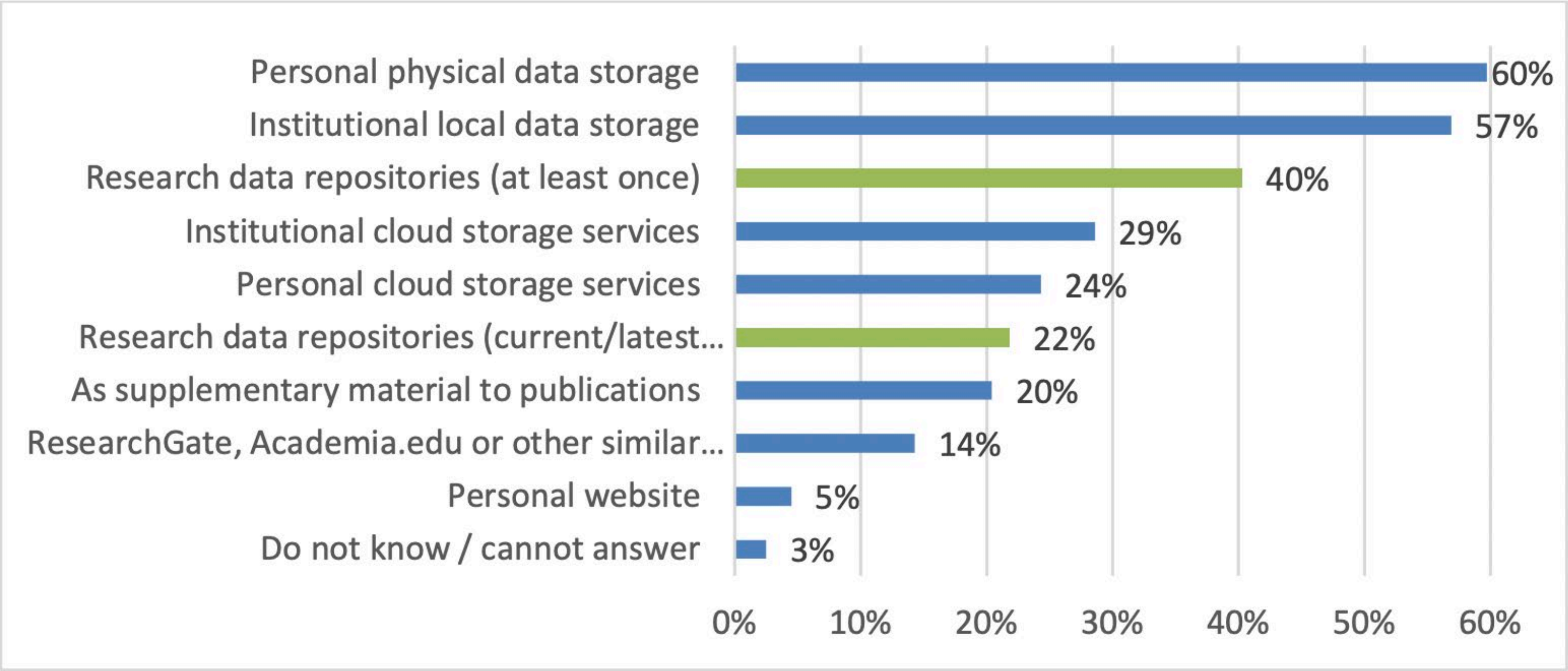
Pujol Priego et al., 2022

Data sharing behaviors



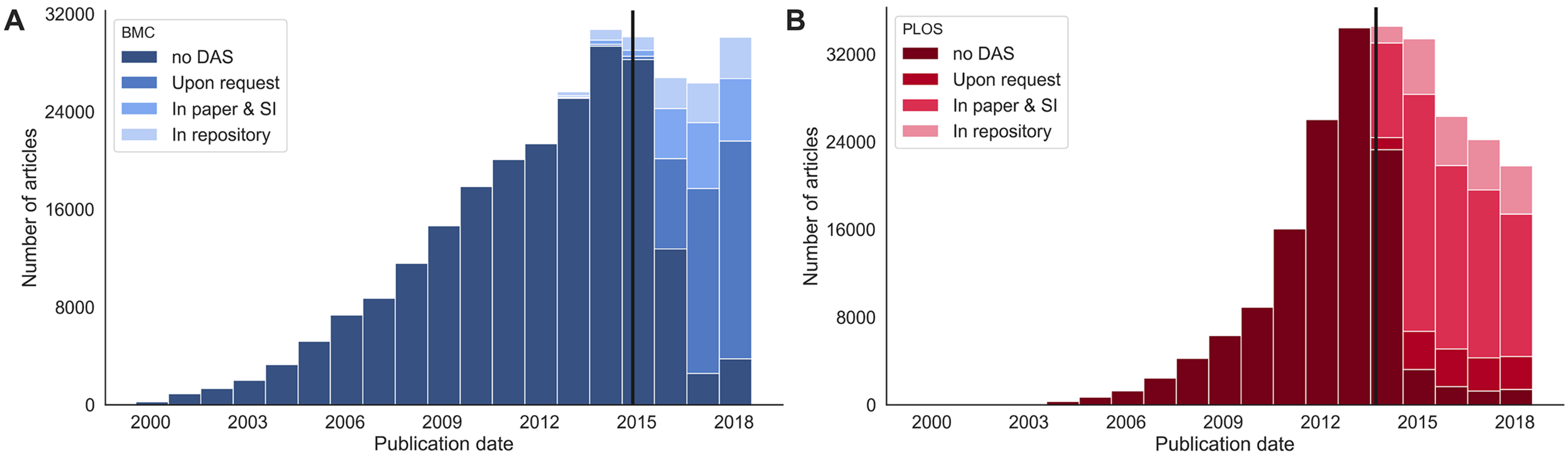
Pujol Priego et al., 2022

Where do researchers store their research data?



European Commission, 2022

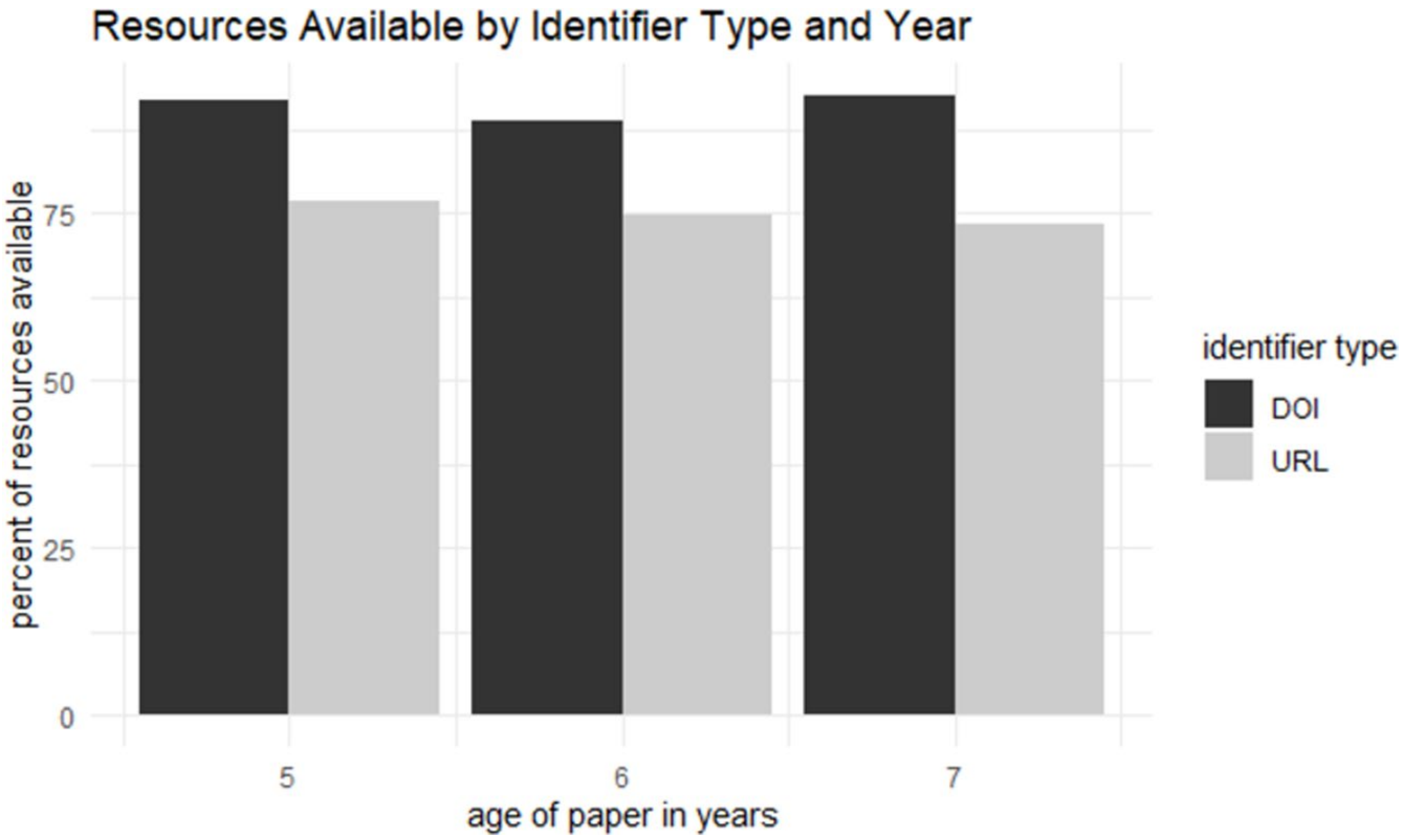
Data Availability Statements Over Time



Correlation of up to 25.36% more citations for articles that share their data in a repository

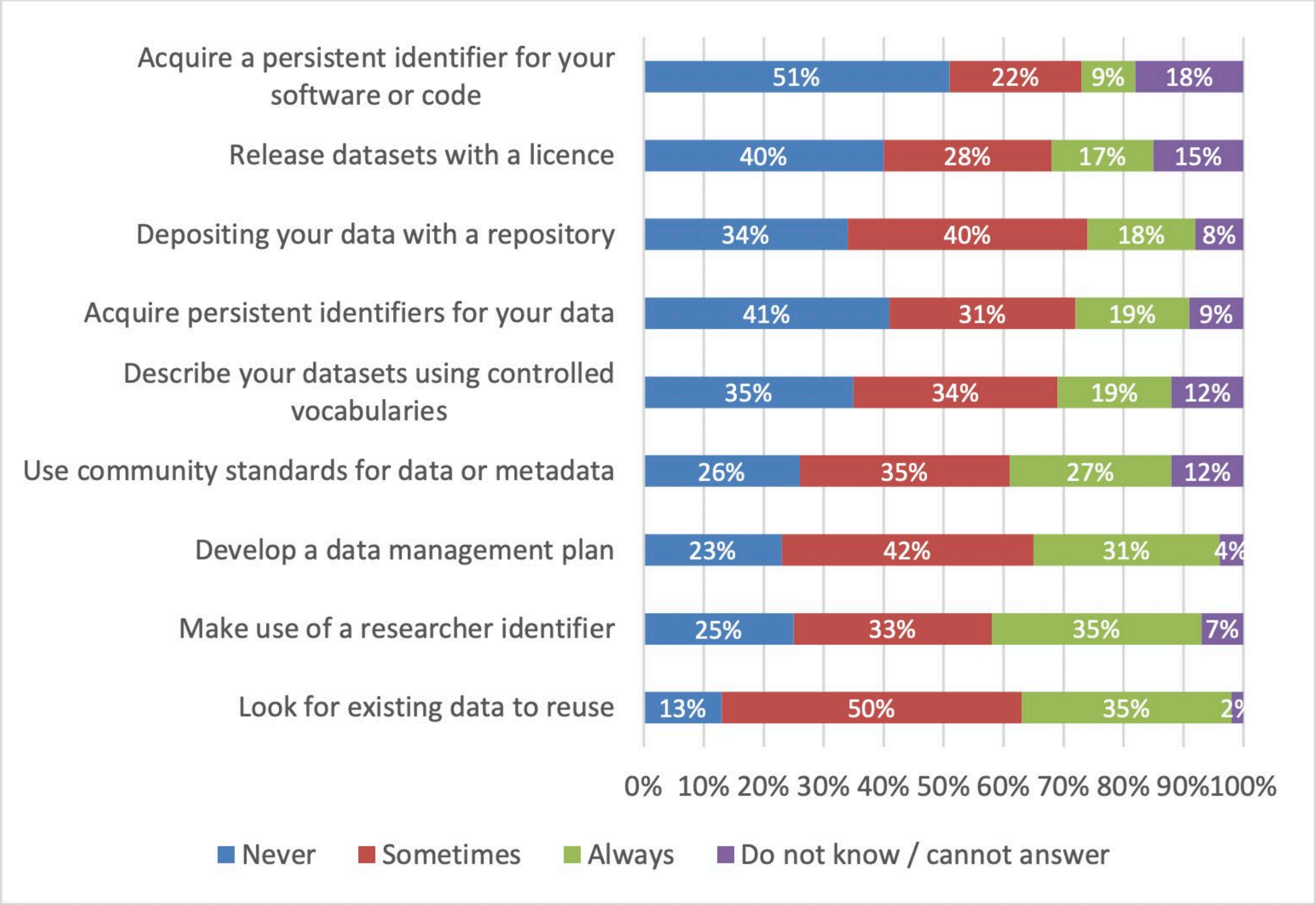
Colavissa et al, 2020

Resource availability with identifier



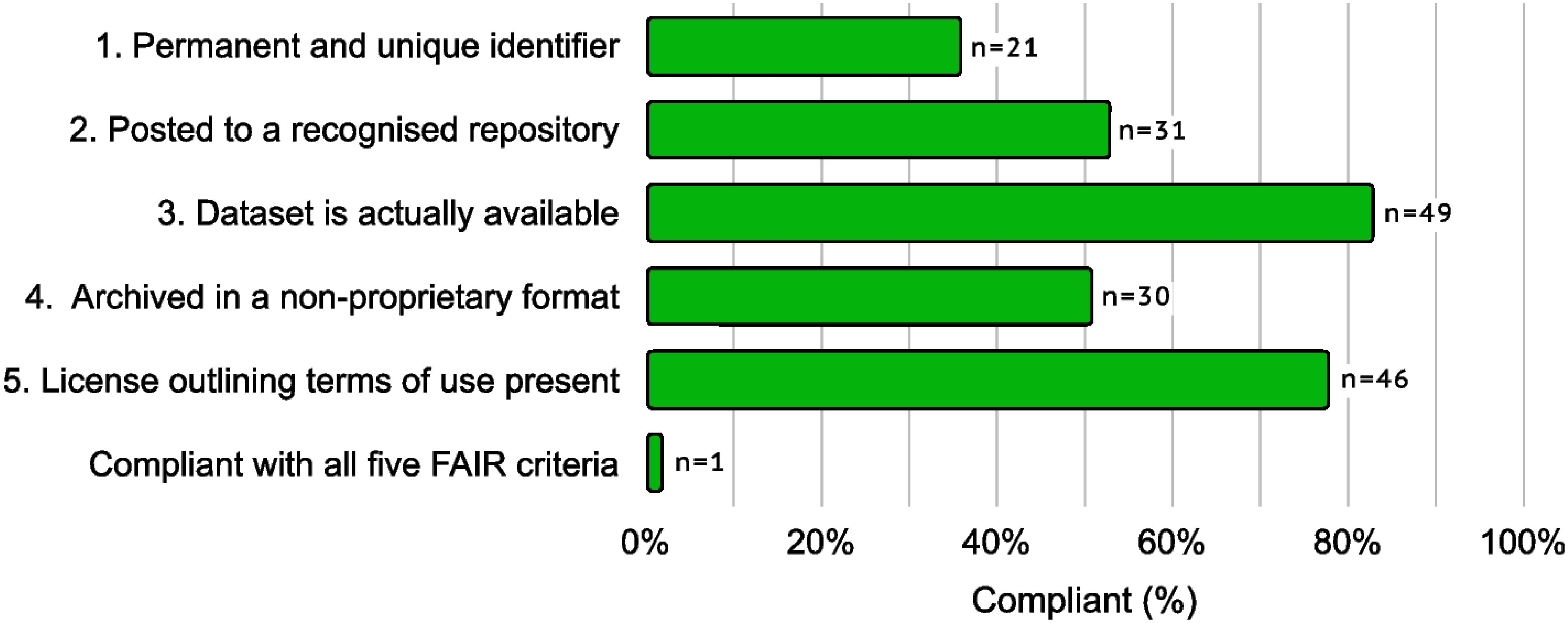
Federer, 2022

Frequency of carrying out specific FAIR-related activities



European Commission, 2022

FAIR assessment of 59 studies



Hamilton et al., 2022

Likely cost of not having FAIR research data

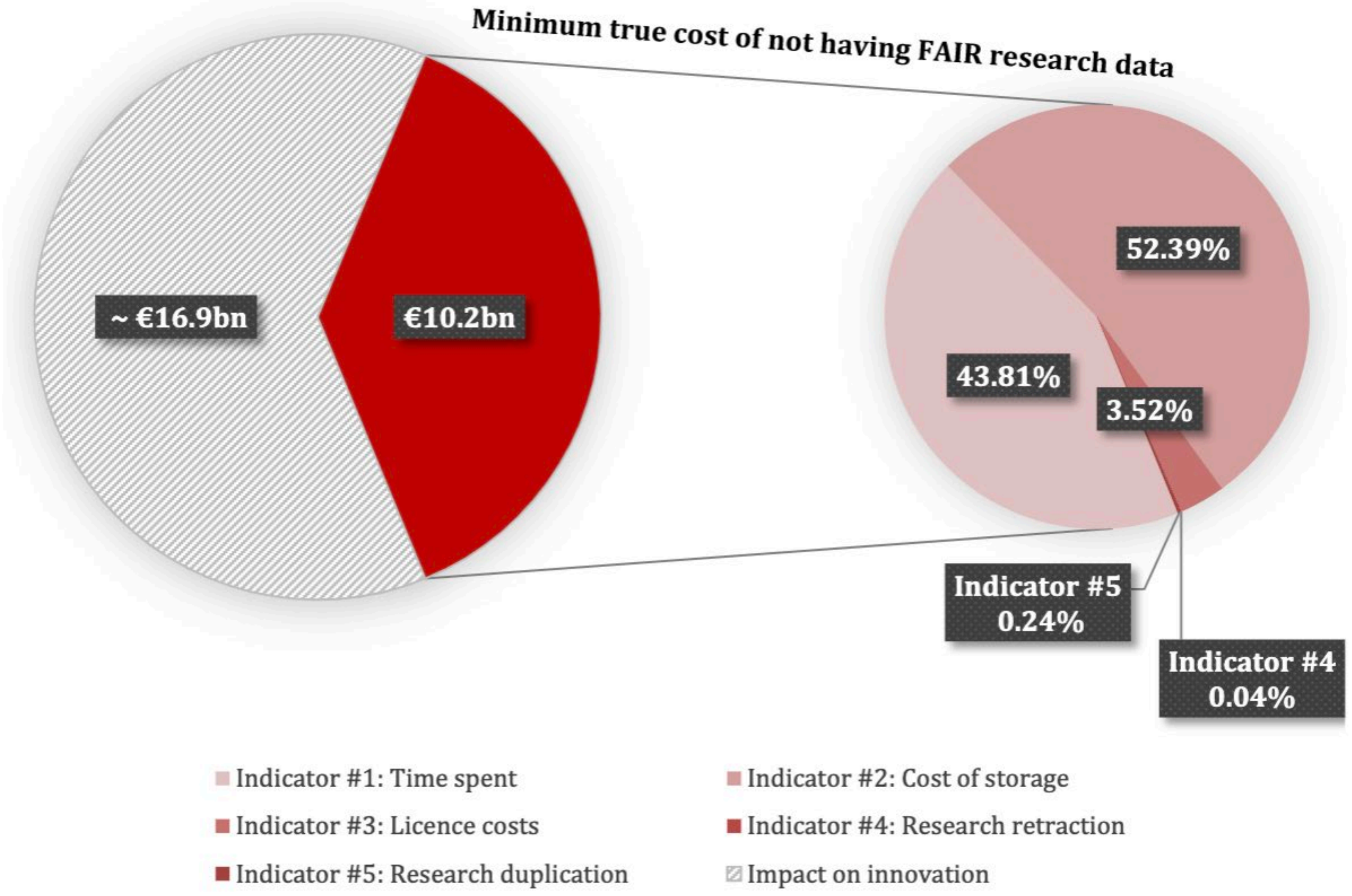
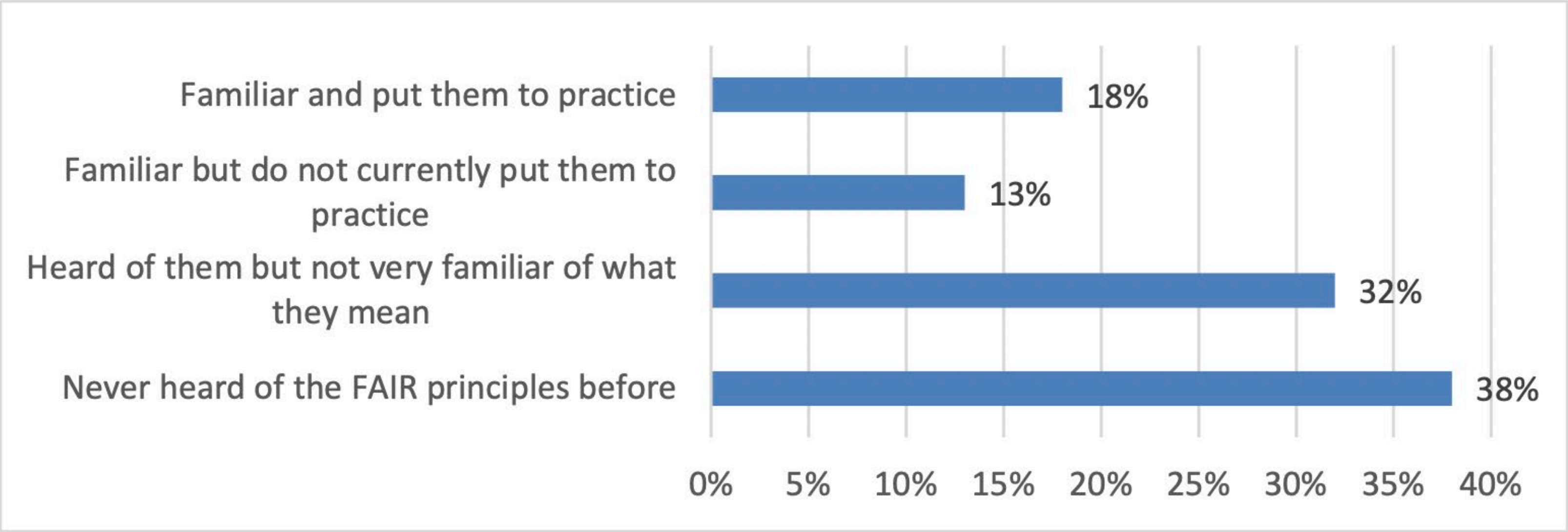


Figure 5: Cost breakdown

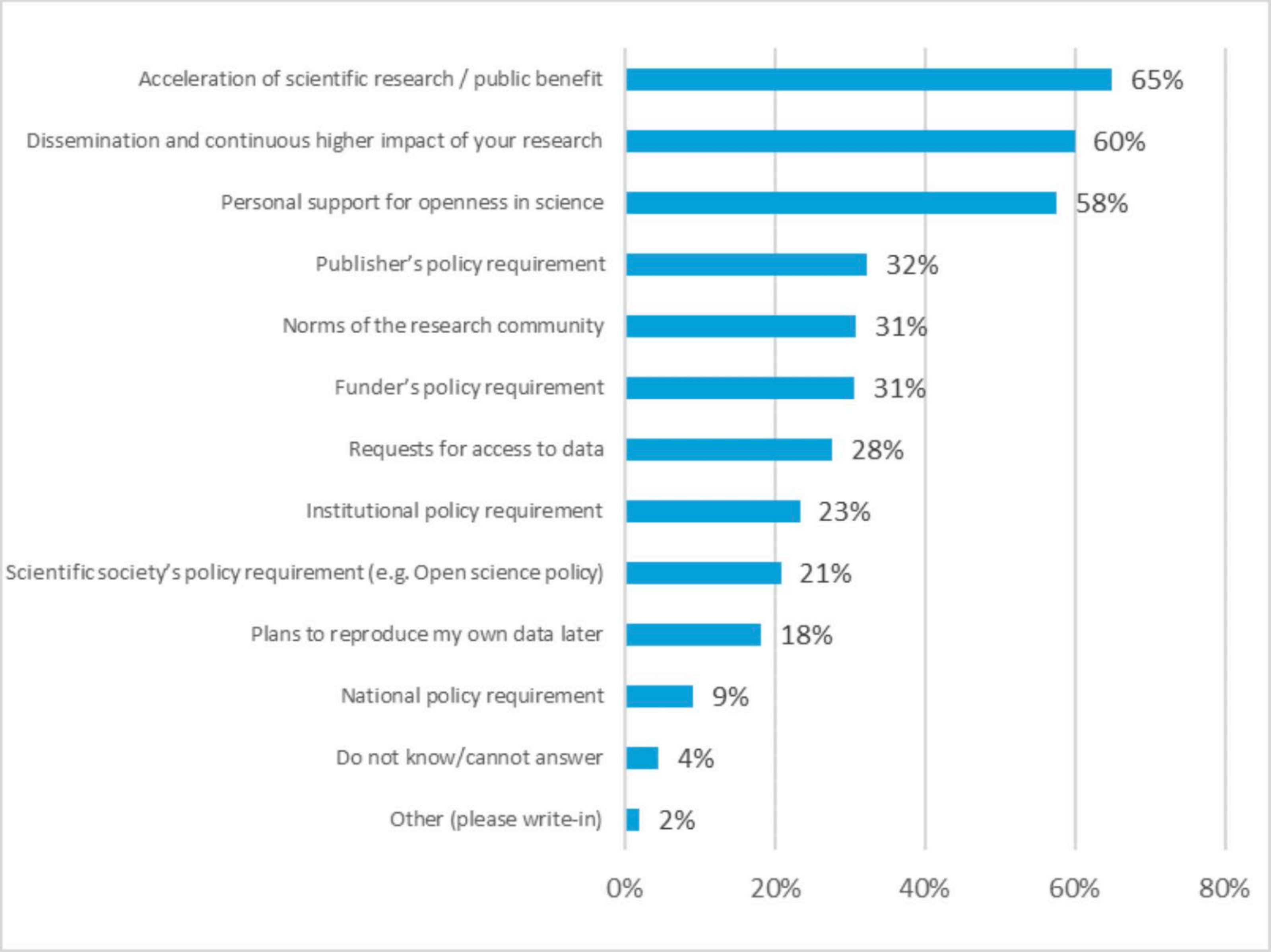
European Commission, 2019

Familiarity with the FAIR principles



European Commission, 2022

Why do researchers store research data in repositories?



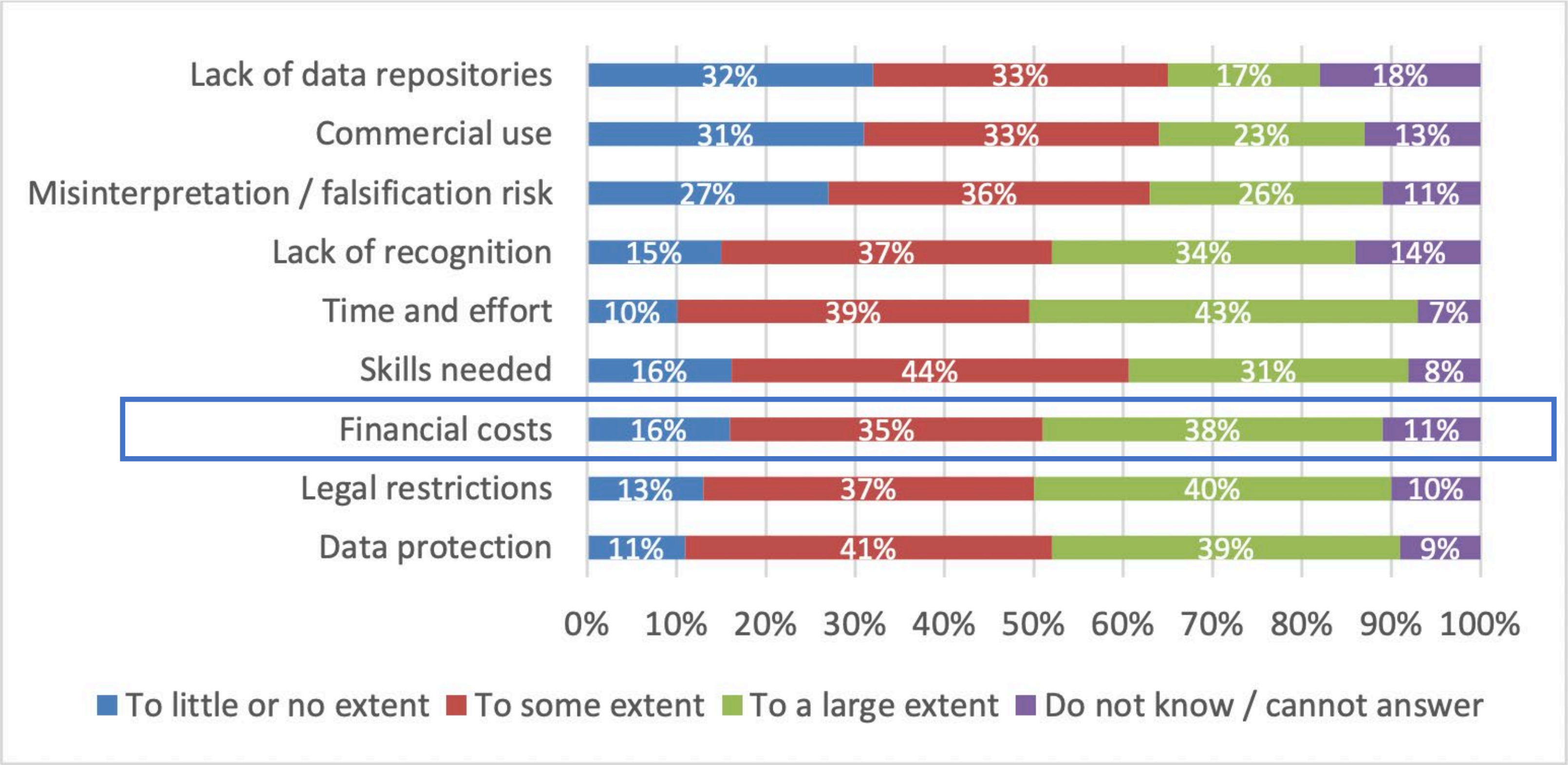
European Commission, 2022

Key barriers

	To a very large extent	To a large extent	To a moderate extent	To a small extent	To a very small extent	Not important in / applicable to my field of research
Pressure to publish for career advancement (N=1,245)	**30%**	**28%**	18%	10%	8%	6%
Lack of overall recognition given to research practices that promote reproducibility (N=1,243)	**20%**	**32%**	22%	11%	8%	8%
Extensive time and effort required to make research reproducible (i.e. describing, sharing, preserving data and methodologies, etc.) (N=1,267)	**16%**	**34%**	28%	10%	8%	5%
Lack of unified guidelines and commonly accepted standards for reproducible research practices (N=1,245)	16%	28%	26%	14%	9%	8%
Insufficient attention is paid to reproducibility-related topics during training and professional development (N=1,246)	15%	28%	29%	13%	8%	6%
Lack of access to the data used or generated by the original research (N=1,239)	17%	26%	23%	15%	11%	7%
Methods require tacit knowledge or particular technical expertise that makes them difficult for others to reproduce (N=1,205)	15%	28%	25%	13%	10%	9%
Focus on reproducibility is not incentivised by home research institutions (e.g. through hiring, tenure, promotion, etc.) (N=1,212)	16%	26%	23%	14%	12%	9%
Lack of journal policies promoting good reproducibility practices (N=1,215)	13%	25%	27%	15%	12%	8%
Research funders do not provide enough incentives to make research reproducible (N=1,218)	13%	23%	25%	15%	16%	9%
Selective reporting of results (including p-hacking / HARKing, lack of reporting of negative / null results) (N=1,058)	11%	25%	23%	15%	10%	16%
Legal or ethical restrictions (e.g. on data sharing) (N=1,264)	16%	19%	19%	14%	16%	16%
Original findings not robust enough (i.e. due to poor research design, statistical analysis, lack of verification or peer-review, etc.) (N=1,200)	10%	23%	28%	17%	13%	9%
Lack of publication of research protocols (N=1,198)	8%	23%	27%	19%	10%	13%
Lack of pre-registration of studies (N=1,058)	5%	15%	21%	20%	15%	24%

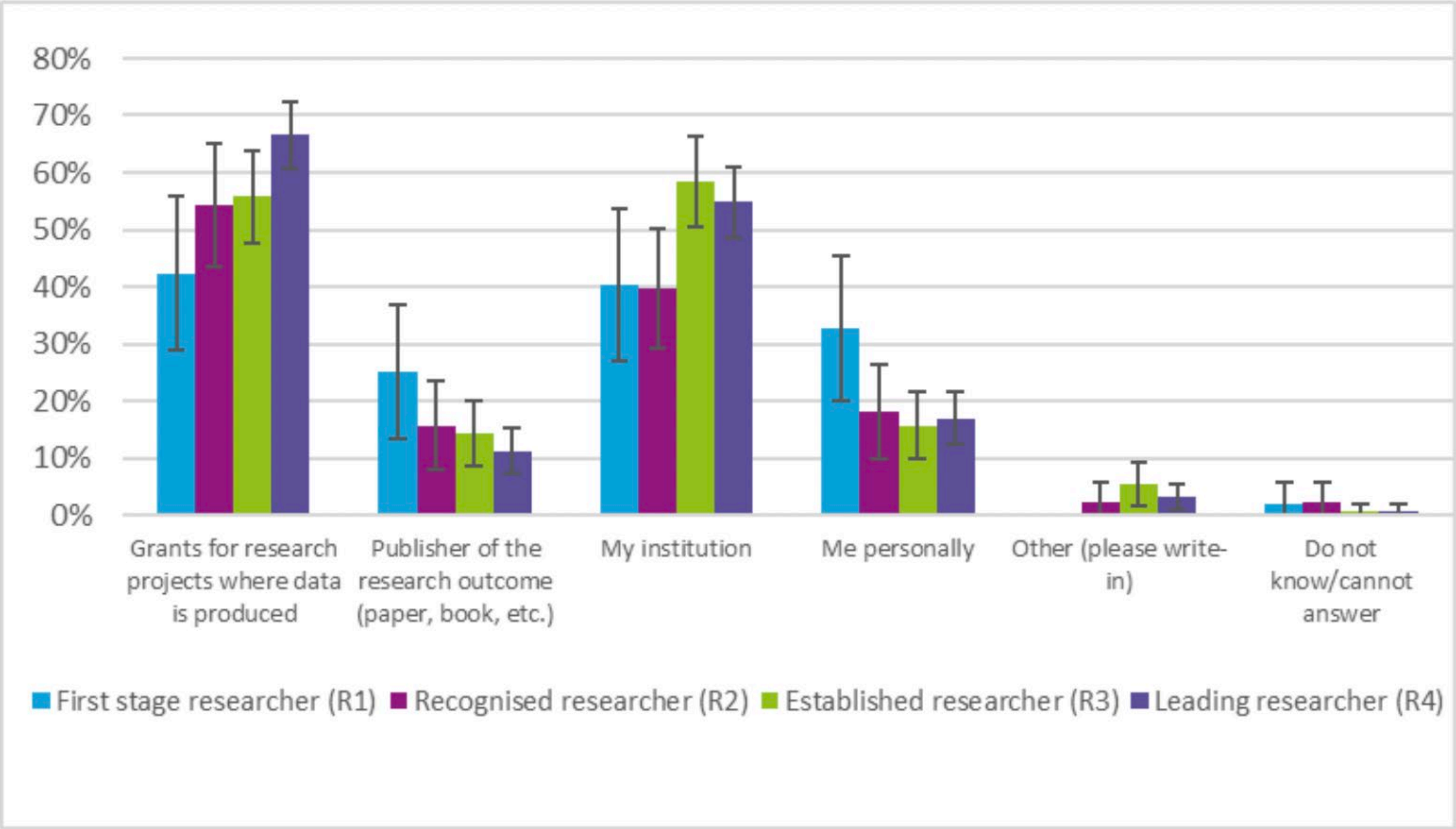
European Commission, 2022

Obstacles to the management and sharing of research data



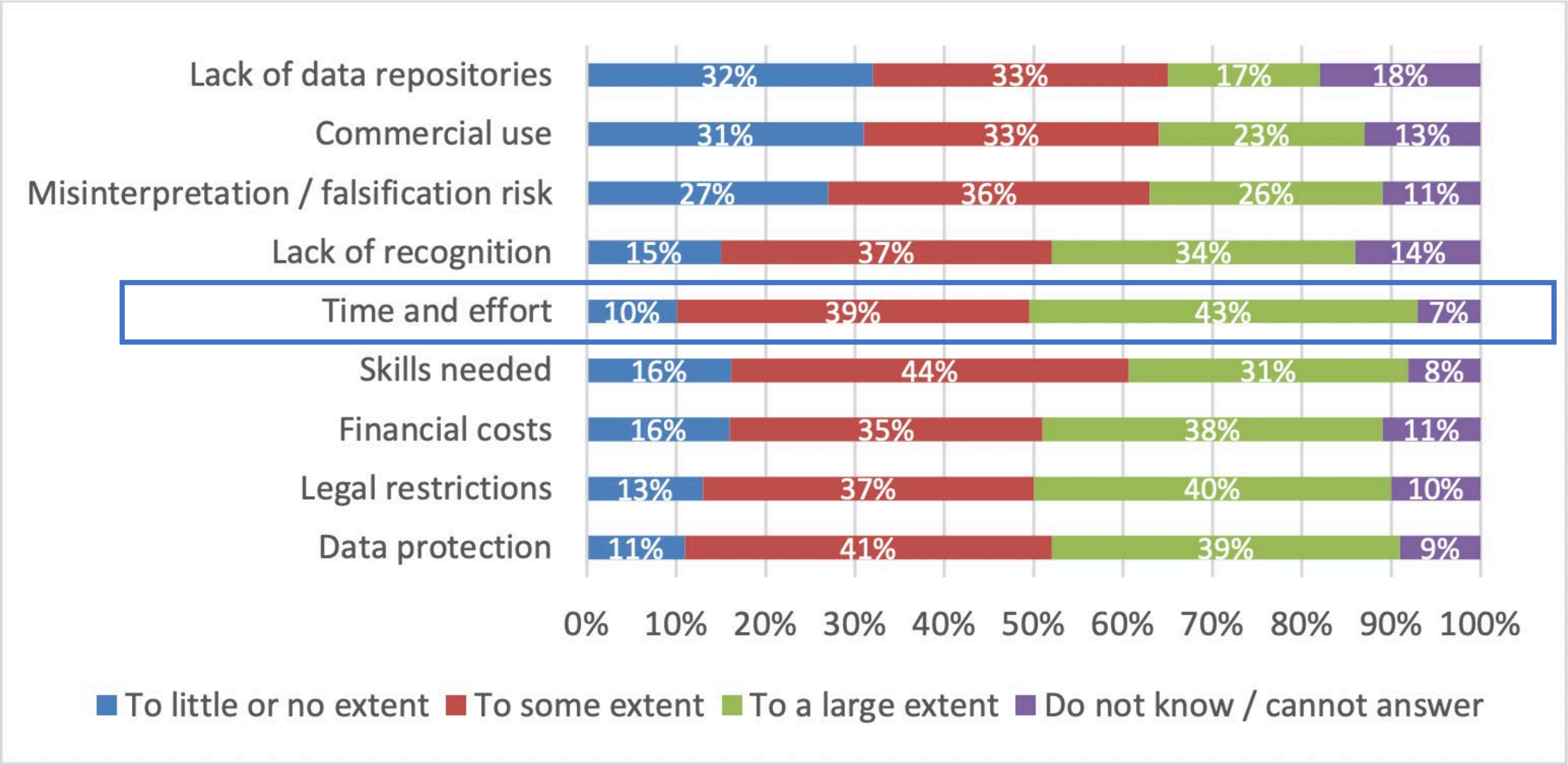
European Commission, 2022

Ways in which research sharing costs were covered



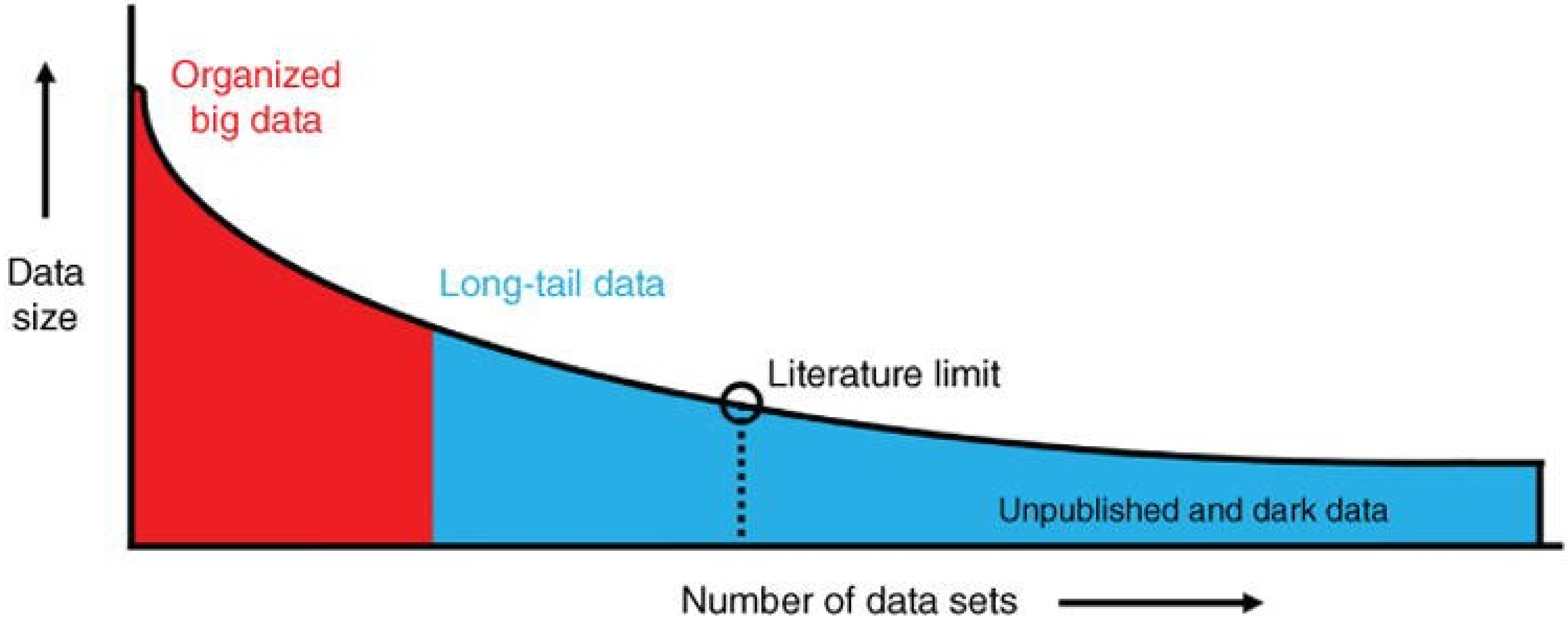
European Commission, 2022

Obstacles to the management and sharing of research data



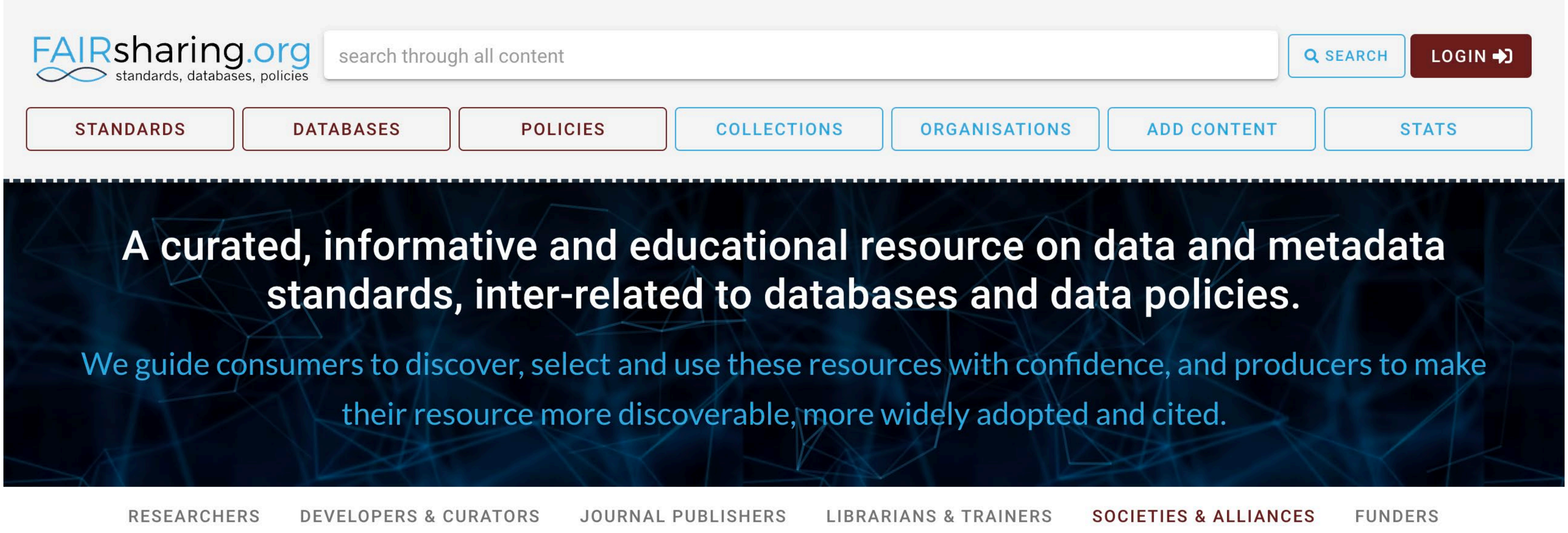
European Commission, 2022

The long tail of data



Ferguson et al., 2014

Many standards



Automate processes?

Automated metadata extraction: challenges and opportunities

Tyler J. Skluzacek
Data Lifecycle and Scalable Workflows Group
Oak Ridge National Laboratory

Kyle Chard and Ian Foster
Department of Computer Science, University of Chicago
Data Science and Learning Division, Argonne National Laboratory

October 07 2022

Automated metadata annotation: What is and is not possible with machine learning

Mingfang Wu, Hans Brandhorst, Maria-Cristina Marinescu, Joaquim More Lopez, Margorie Hlava, Joseph Busch

Check for updates

> Author and Article Information

Data Intelligence 1–17.
https://doi.org/10.1162/dint_a_00162 **Article history**

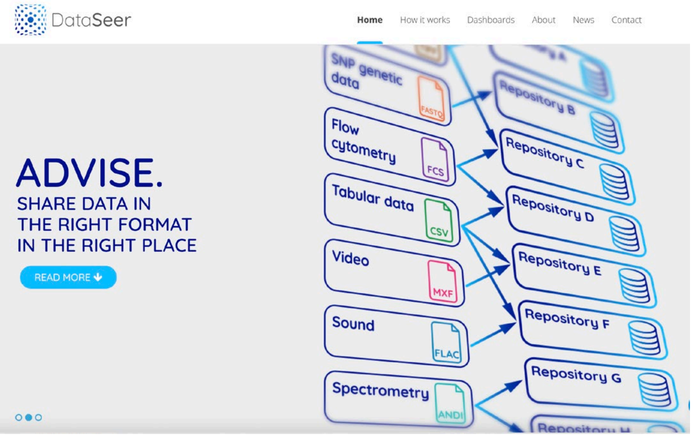
Experience: Automated Prediction of Experimental Metadata from Scientific Publications

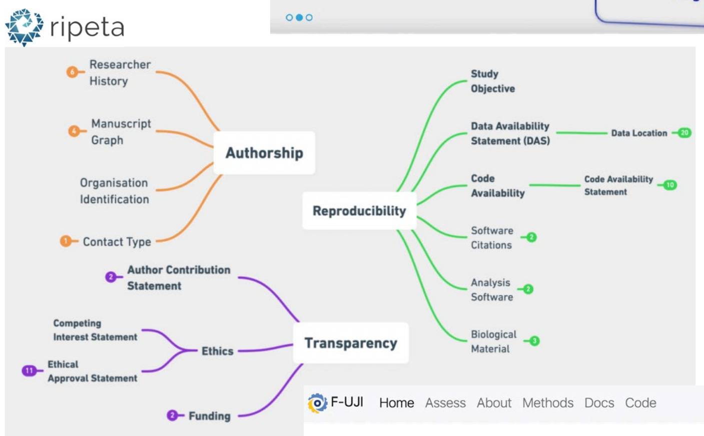
STUTI NAYAK, AMRAPALI ZAVERI, PEDRO HERNANDEZ SERRANO, and MICHEL DUMONTIER, Institute of Data Science, Maastricht University, The Netherlands

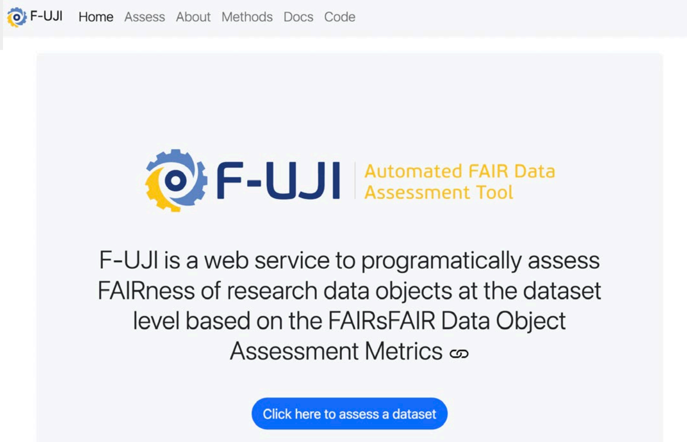
Advancing smart building readiness: Automated metadata extraction using neural language processing methods☆,☆☆

David Waterworth[a,*], Subbu Sethuvenkatraman[b], Quan Z. Sheng[a]

[a] *Department of Computing, Macquarie University, NSW, Australia*
[b] *Energy Business Unit, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia*







Summary

- FAIR data sharing in repositories helps with data transparency, reproducibility, reuse, and impact

- Researchers need help – unaware of FAIR practices and challenges in time, effort, and cost of data sharing

- The ‘long-tail’ of data complicates this further with many options

- Education, support, and workflows/tools to help automate process are potential opportunity areas

References

This slides: https://osf.io/2pdxv

- Errington, T.M., Denis, A., Perfito, N., *et al.* Challenges for assessing replicability in preclinical cancer biology. *eLife*.(2021). https://doi.org/10.7554/eLife.67995
- Gabelica, M., Bojčić, R., & Puljak, L. Many researchers were not compliant with their published data sharing statement: a mixed-meethods study. *J. Clin. Epidemiol.* (2022). https://doi.org/10.1016/j.jclinepi.2022.05.019
- Watson, C. Many researchers say they'll share data – but don't. *Nature News*. (2022). *https://doi.org/10.1038/d41586-022-01692-1*
- Vines, T.H., Albert, A.Y.K., Andrew, R.L., *et al.* The availability of research data declines rapidly with article age. *Curr. Biol*.(2013) https://doi.org/10.1016/j.cub.2013.11.014
- Tedersoo, L., Küngas, R., Oras, E. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci Data* (2021). https://doi.org/10.1038/s41597-021-00981-0
- Ferguson, A.R., Nielson, J.L., Cragin M.H., *et al*. Big data from small data: data-sharing in the 'long-tail' of neuroscience. *Nat. Neurosci.* (2014) https://doi.org/10.1038%2Fnn.3838
- Hamilton, D.G., Page, M.J., Finch, S., *et al.* How often do cancer researchers make their data and code available and what factors are associated with sharing? *BMC Med.* (2022) https://doi.org/10.1186/s12916-022-02644-2
- Pujol Priego, L., Wareham, J., & Romasanta A.K.S. The puzzle of sharing scientific data. *Ind. & Innov.* (2022) https://doi.org/10.1080/13662716.2022.2033178
- Federer, L.M. Long-term availability of data sharing associated with articles in PLOS ONE. *PLOS ONE*. (2022) https://doi.org/10.1371/journal.pone.0272845
- Colavizza, G., Hrynaszkiewicz, I., Staden I., *et al.* The citation advantage of linking publications to research data. *PLOS ONE*. (2020) https://doi.org/10.1371/journal.pone.0230416
- European Commission, Directorate-General for Research and Innovation, Assessing the reproducibility of research results in EU Framework Programmes for Research : final report, *Publications Office of the European Union* (2022) https://data.europa.eu/doi/10.2777/186782
- European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data : cost of not having FAIR research data, *Publications Office of the European Union* (2019) https://data.europa.eu/doi/10.2777/02999
- European Commission, Directorate-General for Research and Innovation, European Research Data Landscape : final report, *Publications Office of the European Union* (2022) https://data.europa.eu/doi/10.2777/3648
- Skluzacek, T., Foster, I., & Chard, K. Automated Metadata Extraction: Challenges and Opportunities. https://www.osti.gov/servlets/purl/1897834
- Wu, M., Brandhorst, H., Marinescu, M-C., *et al*. Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence* (2022) https://doi.org/10.1162/dint_a_00162
- Nayak, S., Zaveri, A., Serrano, P.H., et al. Experience: Automated Prediction of Experimental Metadata from Scientific Publications. *J. Data & Inform Qual*. (2021) https://doi.org/10.1145/3451219
- Waterworth, D., Sethuvenkatraman, S., & Sheng, Q.Z. Advancing smart building readiness: automated metadata extraction using neural language processing methods. *Adv. Appl. Energy* (2021) https://doi.org/10.1016/j.adapen.2021.100041