

# Care to share? Investigating Open Science practices adoption among researchers: a hackathon

This document is for use during the hackathon. It outlines the questions we'll be working on and some key information about the data. If you have any questions please contact Lauren Cadwallader ([lcadwallader@plos.org](mailto:lcadwallader@plos.org)) or Mirela Volaj ([mvolaj@plos.org](mailto:mvolaj@plos.org)).

## Resources:

**Data for the hackathon:** <https://doi.org/10.5281/zenodo.13960084>

**Google Drive to store/share your documents:**  Care to Share - folder for participants

**The original data and documentation:** Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 8). <https://doi.org/10.6084/m9.figshare.21687686>

### Potentially useful code to reuse:

- Grégory Dozot, Kirstie Whitaker, Giovanni Colavizza, & Stéphane Lecorney. (2024). MediaComem/das-public: v2.1 (v2.1). Zenodo. <https://doi.org/10.5281/zenodo.11027260>
  - Related Github repo: <https://github.com/MediaComem/das-public>
- Jean-Quartier, C. (2024). Data on sharing practices of software artifacts and source code for reproducible research (1.0) [Data set]. Graz University of Technology. <https://doi.org/10.3217/gfpp0-2vf87>

## Dataset basics

The PLOS data contains approximately 118,000 PLOS articles published between 2018 and 2024. The Comparator data contains approximately 24,000 articles published in the same time period and on similar topics to the PLOS corpus. The dataset contains the DOIs of each article but information such as title, journal and authors has been removed to discourage ranking of the open science practices.

The data for PLOS and Comparator articles are in different files but they share the same headings so can be merged into one file if preferred. Definitions for the column headings are in the OSI-Column-Descriptions\_v3\_Dec23.pdf file that is included in the hackathon dataset. Data from Dimensions.ai has been added to both data files to provide a list of countries and disciplines associated with each article.

In the dataset we assume all articles generate a preprint but may or may not generate other outputs (data, code). The Study Registration and Protocols sharing indicators have been added to the data files (they are separate in the original dataset), however, as these indicators are still in development they have not been run for all the articles in the dataset. Study Registration has not been run for articles published in 2024. The Protocols indicator was only run for PLOS

articles published from January 2019 to the end of June 2023 and for a sub-sample of the Comparator set.

## Question 1: Do researchers from different **countries** exhibit different open science behaviours?

As well as thinking about questions like “What is the data sharing rate in The Netherlands compared to the USA?” we can go further and ask questions like, “Do researchers in The Netherlands share all the open science outputs that they generate and how does this compare to other countries?”

Useful variables in the dataset

Variable name	What is it?	Data format	Notes
Data_Generated	Indicates if the article generated data as part of the reported research	Yes / No	
Data_Shared	Indicates if the article shared data related to the article	Yes / No	For data to be deemed as “shared” it must be available online (with URL, DOI or Accession ID or in Supplementary Information). Reused data will also be identified as shared.
Preprint_Match	Indicates if the article shared a preprint	Yes / No	
Code_Generated	Indicates if the article generated code as part of the reported research	Yes / No	
Code_Shared	Indicates if the article shared code related to the article	Yes / No	For code to be deemed as “shared” it must be available online (with URL, DOI or Accession ID or in Supplementary Information).
Dimensions_Country	Lists all the countries associated with the authors following the authorship order	Country 1; Country 2; Country 3 etc	Country names are repeated if multiple authors are from the same country

Other useful variables might be:

- *Protocol\_Shared* and *Registration\_Shared* if you want to include these open science practices although this information is not available for all articles.
- *Publication\_Year* or *Quarter* if you want to look at this over time.
- You could look at where data is shared and think about whether the researchers are working towards making their data FAIR. For this you might want to use:
  - *Repositories\_data* will list the name of a repository if one was used and it is included on the OSI repository list of 139 commonly used repositories.
  - *URL\_data* will give the URL of any shared data including data shared in a repository. It is more inclusive than *Repositories\_data* but will also include things like lab websites.
  - *Data\_Location* indicates whether shared data is “Online”, in “Supplementary Information” or in both. Values are comma separated.

## Question 2: Do researchers from different **disciplines** exhibit different open science behaviours?

As well as thinking about questions like “What is the data sharing rate in the Medical Sciences compared to the Plant Sciences?” we can go further and ask questions like, “Do researchers in the Medical Science share all the open science outputs that they generate and how does this compare to other disciplines?”

Useful variables in the dataset

Variable name	What is it?	Data format	Notes
Data_Generated	Indicates if the article generated data as part of the reported research	Yes / No	
Data_Shared	Indicates if the article shared data related to the article	Yes / No	For data to be deemed as “shared” it must be available online (with URL, DOI or Accession ID or in Supplementary Information). Reused data will also be identified as shared.
Preprint_Match	Indicates if the article shared a preprint	Yes / No	
Code_Generated	Indicates if the article generated code as part of the reported research	Yes / No	
Code_Shared	Indicates if the article shared code related to the article	Yes / No	For code to be deemed as “shared” it must be available online (with URL, DOI or

			Accession ID or in Supplementary Information).
Dimensions_FoR	Indicates the disciplines the article falls into	Uses Fields of Research hierarchical categories ( <a href="#">scheme</a> ; <a href="#">background</a> ; <a href="#">hierarchy</a> <a href="#">breakdown</a> - see Table 1)	We suggest you focus on the top level categories which have a two digit prefix. There are 23 different disciplines in this level. Articles can have more than one discipline assigned even at the top level of the hierarchy.

Other useful variables might be:

- *Protocol\_Shared* and *Registration\_Shared* if you want to include these open science practices although this information is not available for all articles.
- *Publication\_Year* or *Quarter* if you want to look at this over time.
- You could look at where data is shared and think about whether the researchers are working towards making their data FAIR. For this you might want to use:
  - *Repositories\_data* will list the name of a repository if one was used and it is included on the OSI repository list of 139 commonly used repositories.
  - *URL\_data* will give the URL of any shared data including data shared in a repository. It is more inclusive than *Repositories\_data* but will also include things like lab websites.
  - *Data\_Location* indicates whether shared data is “Online”, in “Supplementary Information” or in both. Values are comma separated.

## Extension question (if there is time or the desire)

Do open science behaviours have an impact on when a preprint is shared? Do researchers share their preprints earlier if they also share their other open science outputs? The dataset contains the date the preprint was published as well as the date the article was published.