

# README for PLOS Open Science Indicators Dataset, version 8

This readme file was generated on 22-09-2024 by Lauren Cadwallader.

## GENERAL INFORMATION

Title of Dataset: PLOS Open Science Indicators

### Author Information

Name: Public Library of Science

Recommended citation for this dataset:

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

### Named contact Information

Name: Iain Hrynaskiewicz

ORCID: 0000-0002-9673-5559

Institution: Public Library of Science

Email: [ihrynaskiewicz@plos.org](mailto:ihrynaskiewicz@plos.org) / [plos@plos.org](mailto:plos@plos.org)

### Alternate Contact Information

Name: Lauren Cadwallader

ORCID: 0000-0002-7571-3502

Institution: Public Library of Science

Email: [lcadwallader@plos.org](mailto:lcadwallader@plos.org) / [plos@plos.org](mailto:plos@plos.org)

Version: 8

Date of data collection: Comparator-Dataset\_v1\_Dec22.csv XML was collected 15-09-2022. PLOS-Dataset\_v1\_Dec22.csv XML was collected 03-08-2022. Additional data for the version 2 dataset was collected 02-23-2023 (PLOS-Dataset\_v2\_Mar23.csv) and 14-03-2023 (Comparator-Dataset\_v2\_Mar23.csv). Additional data for the version 3 dataset was collected 11-04-2023. Additional data for the version 4 dataset was collected 10-7-2023 (PLOS-Dataset\_v4\_Sep23.csv) and 14-7-2023 (Comparator-Dataset\_v4\_Sep23.csv). Additional data for the version 5 dataset was collected 02-11-2023 (PLOS-Dataset\_v5\_Dec23.csv) and 02-11-2023 (Comparator-Dataset\_v5\_Dec23.csv). Additional data for the version 6 dataset was collected 15-01-2024 (PLOS-Dataset\_v6\_Mar24.csv) and 16-01-2024 (Comparator-Dataset\_v6\_Mar24.csv). Additional data for the version 7 dataset was collected 15-04-2024 (PLOS-Dataset\_v7\_Jun24.csv) and 16-04-2024

(Comparator-Dataset\_v7\_Jun24.csv). Additional data for the version 8 dataset was collected 15-07-2024 (PLOS-Dataset\_v8\_Sep24.csv) and 08-08-2024 (Comparator-Dataset\_v8\_Sep24.csv).

Information about funding sources that supported the collection of the data: No external funding was received for this work.

## SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data: CC BY 4.0

Links to publications that cite or use the data:

<https://theplosblog.plos.org/2022/12/open-science-indicators-first-dataset> (cites v1 data)  
<https://theplosblog.plos.org/2023/04/open-science-indicators/> (cites v2 data)  
<https://theplosblog.plos.org/2023/06/Open-Science-Indicators-Update-Q1-2023> (cites v3 data)  
<https://theplosblog.plos.org/2023/10/open-science-indicators-q2-2023> (cites v4 data)  
<https://theplosblog.plos.org/2023/10/measuring-protocol-sharing> (cites v4 data)  
<https://theplosblog.plos.org/2024/01/open-science-indicators-q3-23/> (cites v5 data)  
<https://theplosblog.plos.org/2024/03/six-years-of-open-science-indicators-data/> (cites v6 data)  
<https://theplosblog.plos.org/2024/07/a-new-open-science-indicator-measuring-study-registration/> (cites v7 data)

Was data derived from another source?

If yes, list source(s):

Yes. Data was partly derived from journal article metadata accessed via PubMed Central or the 'all of PLOS' API (<https://github.com/PLOS/allofplos>). Full details are in the OSI-Methods-Statement\_v8\_Sep24.pdf file.

## UPDATES ON PREVIOUS VERSION OF THE DATASET

- An additional 5,597 articles with publication dates between 1 April 2024 and 30 June 2024 have been added to the PLOS-Dataset\_v8\_Sep24.csv.
- The Comparator-Dataset\_v8\_Sep24.csv has had an additional 1,120 articles added, which were published between 1 April 2024 and 30 June 2024.
- The Comparator sample for articles published between 1 January 2024 and 31 March 2024 has been resampled due to issues with the sampling methodology that was detected after the release of v7. Details of the resampling method can be found in OSI-Methods-Statement\_v8\_Sep24.pdf.
- The "Preliminary Release for Protocols Indicator" folder contains the same files shared in v4. The "Preliminary Release for Study Registration" folder contains the same files shared in v7,
- A very small number of articles (<100) could not be processed by the algorithm despite repeated runs and "NA" is present for all of the columns

## DATA & FILE OVERVIEW

### Folder: Main Data Files

#### PLOS-Dataset\_v8\_Sep24.csv

Data pertaining to the PLOS corpus of articles. Contains information extracted from PLOS article XML and Open Science Indicators information as described in the OSI-Column-Descriptions\_v3\_Dec23.pdf file.

#### Comparator-Dataset\_v8\_Sep24.csv

Data pertaining to the comparator corpus of articles. Contains article xml and Open Science Indicators information as described in the OSI-Column-Descriptions\_v3\_Dec23.pdf file. Inclusion criteria for this set is described in the OSI-Methods-Statement\_v8\_Sep24.pdf file.

#### OSI-Summary-statistics\_v8\_Sep24.xlsx

This file contains the summary data (PLOS and comparator). The data are arranged in tabular form and present summary statistics for each of the Open Science Indicators (data, code and preprints).

### Folder: Main Documentation

#### OSI-Methods-Statement\_v8\_Sep24.pdf

Text document describing the methods underlying the data collection and analysis. It includes details of how the PLOS and comparator articles were selected and accessed, methods behind data and code generation and sharing detection, preprint detection and accuracy rates for the analysis. Methods and accuracy rates will be developed and improved in future releases.

#### OSI-Column-Descriptions\_v3\_Dec23.pdf

Text document describing the fields used in PLOS-Dataset\_v8\_Sep24.csv and Comparator-Dataset\_v8\_Sep24.csv.

#### OSI-Repository-List\_v1\_Dec22.xlsx

List of repositories and their characteristics used to identify specific repositories in the PLOS-Dataset\_v8\_Sep24.csv and Comparator-Dataset\_v8\_Sep24.csv repository fields. The list is derived from commonly used repositories. It is not designed to be an exhaustive list of all possible repositories and will be developed over time.

### Folder: Preliminary Release for Protocols Indicator

#### Protocols-Dataset\_Sep23.csv

Data on protocol sharing pertaining to the PLOS and Comparator corpus of articles. Contains information extracted and analysed as described in Protocols-Methods-Statement\_Sep23.pdf. Column headings are described in Protocols-Column-Headings\_Sep23.pdf.

Protocols-Summary-Statistics\_Sep23.xlsx

This file contains the summary data (PLOS and comparator), some of which is used within the related blog post <https://theplosblog.plos.org/2023/10/measuring-protocol-sharing>. The data are arranged in tabular form.

Protocols-Methods-Statement\_Sep23.pdf

Text document describing the methods underlying the data collection and analysis for the protocols indicator. It includes details of how the PLOS and comparator articles were accessed, rationale and methods behind protocol detection, and accuracy rates for the analysis.

Protocols-Column-Headings\_Sep23.pdf

Text document describing the fields used in Protocols-Dataset\_Sep23.csv.

### **Folder: Preliminary Release for Study Registration Indicator**

Study-Registration-Dataset\_Jun24.csv

Data on study registration sharing pertaining to the PLOS and Comparator corpus of articles. Contains information extracted and analysed as described in Study-Registration-Methods-Statement\_Jun24.pdf. Column headings are described in Registration-Column-Headings\_Jun24.pdf.

Study-Registration-Summary-Statistics\_Jun24.xlsx

This file contains the summary data for Study Registration (PLOS and comparator). The data are arranged in tabular form.

Study-Registration-Methods-Statement\_Jun24.pdf

Text document describing the methods underlying the data collection and analysis for the study registration indicator. It includes the definition of study registration, methods behind study registration detection and accuracy rates for the analysis.

Registration-Column-Headings\_Jun24.pdf.

Text document describing the fields used in Study-Registration-Dataset\_Jun24.csv.

OSI-Registration-Databases-List\_Jun24.xlsx

List of registration databases and their characteristics used to identify specific registrations in the Study-Registration-Dataset\_Jun24.csv. The registries included in the list cover registration of 4 different types of study: clinical trials, systematic reviews, animal studies, and other types of study (or general purpose registries).

### **METHODOLOGICAL INFORMATION**

Description of methods used for collection/generation of data:

For a description of the methods used to collect/generate the data see

OSI-Methods-Statement\_v8\_Sep24.pdf.

Methods for processing the data:

Summary statistics were generated using Tableau and Excel.

Describe any quality-assurance procedures performed on the data:

For information on accuracy rates and how these were assessed see OSI-Methods-Statement\_v8\_Sep24.pdf.

People involved with sample collection, processing, analysis and/or submission:

Allegra Pearce, Tim Vines, Asura Enkhbayar, Michael Shinkoda, Scott Kerr of DataSeer (<https://dataseer.ai/>) for data acquisition and supporting information. Lauren Cadwallader of PLOS for data processing and supporting information.

DATA-SPECIFIC INFORMATION FOR: PLOS-Dataset\_v8\_Sep24.csv

Number of variables: 27

Number of cases/rows: 117,827

Variable List: See OSI-Column-Descriptions\_v3\_Dec23.pdf

Missing data codes: “Missing” for “Discipline” field. “NA” is used where the field is not relevant.

Specialized formats or other abbreviations used:

- “DA” = Data Availability Statement

DATA-SPECIFIC INFORMATION FOR: Comparator-Dataset\_v8\_Sep24.csv

Number of variables: 27

Number of cases/rows: 24,106

Variable List: See OSI-Column-Descriptions\_v3\_Dec23.pdf

Missing data codes: “Missing” for “Country” field. “NA” is used where the field is not relevant and for “DA” field if no statements have been detected.

Specialized formats or other abbreviations used:

- “DA” = Data Availability Statement

DATA-SPECIFIC INFORMATION FOR: Protocols-Dataset\_Sep23.csv

Number of variables: 8

Number of cases/rows: 96363

Variable List: See Protocols-Column-Headings\_Sep23.pdf

Missing data codes: "N/A" is used where the field is not relevant or missing

Specialized formats or other abbreviations used: None

DATA-SPECIFIC INFORMATION FOR: Study-Registration-Dataset\_Jun24.csv

Number of variables: 11

Number of cases/rows: 129802

Variable List: See Registration-Column-Headings\_Jun24.pdf.

Missing data codes: Cells are blank where the field is not relevant or missing

Specialized formats or other abbreviations used: Registry names are abbreviated. See OSI-Registration-Databases-List\_Jun24.xlsx for information.