

# README for PLOS Open Science Indicators Dataset, version 4

This readme file was generated on 21-9-2023 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 4). <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: 4

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).

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)

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\_v4\_Sep23.pdf file.

## UPDATES ON PREVIOUS VERSION OF THE DATASET

- An additional 4,233 articles have been to PLOS-Dataset\_v4\_Sep23.csv. These all have publication dates between 1 April 2023 and 30 June 2023.
- The Comparator-Dataset\_v4\_Sep23.csv has been extended across all year and now includes publications to 30 June 2023. An additional 7,847 articles have been added.
- Additional guidance on the detection of code generation has been added to the OSI-Methods-Statement\_v4\_Sep23.pdf file as the algorithm that detects code generation was updated between the creation of versions 1 and 2 of the Open Science Indicators datasets. Caution is needed when interpreting code generation results for 2022 and 2023 articles in this version of the data, but this issue will be resolved in the planned version 5 release of the data (December 2023).
- Preliminary data for a 4th open science indicator - protocol sharing - has been released separately to the main dataset along with supporting documentation.
- Dataset files have been organised into folders.

## DATA & FILE OVERVIEW

Folder: Main Data Files

PLOS-Dataset\_v4\_Sep23.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\_v2\_Mar23.pdf file.

Comparator-Dataset\_v4\_Sep23.csv

Data pertaining to the comparator corpus of articles. Contains article xml and Open Science Indicators information as described in the OSI-Column-Descriptions\_v2\_Mar23.pdf file. Inclusion criteria for this set is described in the OSI-Methods-Statement\_v4\_Sep23.pdf file.

OSI-Summary-statistics\_v4\_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/open-science-indicators-q2-2023> . 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\_v4\_Sep23.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\_v2\_Mar23.pdf

Text document describing the fields used in PLOS-Dataset\_v4\_Sep23.csv and Comparator-Dataset\_v4\_Sep23.csv.

OSI-Repository-List\_v1\_Dec22.xlsx

List of repositories and their characteristics used to identify specific repositories in the PLOS-Dataset\_v4\_Sep23.csv and Comparator-Dataset\_v4\_Sep23.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.

## 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\_v4\_Sep23.pdf.

Methods for processing the data:

Summary statistics were generated using Microsoft Excel.

Describe any quality-assurance procedures performed on the data:

For information on accuracy rates and how these were assessed see OSI-Methods-Statement\_v4\_Sep23.pdf.

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

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

## DATA-SPECIFIC INFORMATION FOR: PLOS-Dataset\_v4\_Sep23.csv

Number of variables: 25

Number of cases/rows: 78363

Variable List: See OSI-Column-Descriptions\_v2\_Mar23.pdf

Missing data codes: "Missing" for "Discipline" field. "N/A" is used where the field is not relevant.

Specialized formats or other abbreviations used:

- "DA" = Data Availability Statement

## DATA-SPECIFIC INFORMATION FOR: Comparator-Dataset\_v4\_Sep23.csv

Number of variables: 25

Number of cases/rows: 16023

Variable List: See OSI-Column-Descriptions\_v2\_Mar23.pdf

Missing data codes: "Missing" for "Country" field. "N/A" 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