

README for PLOS Open Science Indicators Dataset, version 1

This readme file was generated on 7-12-2022 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 1).
<https://doi.org/10.6084/m9.figshare.21687686>.

Named contact Information

Name: Iain Hrynaszkiewicz

ORCID: 0000-0002-9673-5559

Institution: Public Library of Science

Email: ihrynaszkiewicz@plos.org / plos@plos.org

Alternate Contact Information

Name: Lauren Cadwallader

ORCID: 0000-0002-7571-3502

Institution: Public Library of Science

Email: lcadwallader@plos.org / plos@plos.org

Version: 1

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.

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>

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_v1_Dec22.pdf file.

DATA & FILE OVERVIEW

File List:

PLOS-Dataset_v1_Dec22.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_v1_Dec22.pdf file.

Comparator-Dataset_v1_Dec22.csv

Data pertaining to the comparator corpus of articles. Contains article xml and Open Science Indicators information as described in the OSI-Column-Descriptions_v1_Dec22.pdf file. Inclusion criteria for this set is described in the OSI-Methods-Statement_v1_Dec22.pdf file.

OSI-Summary-statistics_v1_Dec22.xlsx

This file contains the summary data (PLOS and comparator), some of which is used within the related blog post <https://theplosblog.plos.org/2022/12/open-science-indicators-first-dataset>. The data are arranged in tabular form and present summary statistics for each of the Open Science Indicators (data, code and preprints).

OSI-Methods-Statement_v1_Dec22.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_v1_Dec22.pdf

Text document describing the fields used in PLOS-Dataset_v1_Dec22.csv and Comparator-Dataset_v1_Dec22.csv.

OSI-Repository-List_v1_Dec22.xlsx

List of repositories and their characteristics used to identify specific repositories in the PLOS-Dataset_v1_Dec22.csv and Comparator-Dataset_v1_Dec22.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.

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_v1_Dec22.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_v1_Dec22.pdf.

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

Allegra Pearce and Tim Vines 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_v1_Dec22.csv

Number of variables: 22

Number of cases/rows: 61318

Variable List: See OSI-Column-Descriptions_v1_Dec22.pdf

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

Specialized formats or other abbreviations used:

- Dates are in the format dd/mm/yyyy.
- “DA” = Data Availability Statement

DATA-SPECIFIC INFORMATION FOR: Comparator-Dataset_v1_Dec22.csv

Number of variables: 22

Number of cases/rows: 6588

Variable List: See OSI-Column-Descriptions_v1_Dec22.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:

- Dates are in the format dd/mm/yyyy.
- "DA" = Data Availability Statement