

README for PLOS Open Science Indicators Dataset, version 3

This readme file was generated on 21-6-2023 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 3). <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 / 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: 3

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.

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)

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_v3_Jun23.pdf file.

UPDATES ON PREVIOUS VERSION OF THE DATASET

- 1595 duplicate entries that were present in the v2 data have been removed from the PLOS dataset (PLOS-Dataset_v3_Jun23.csv). The duplicates were confined to the Publication Months July and August 2022 in the v2 data.
- An additional 4616 articles have been to PLOS-Dataset_v3_Jun23.csv and 541 to Comparator-Dataset_v3_Jun23.csv. These all have publication dates between 1 Jan 2023 and 31 March 2023.
- OSI-Dimensions-FoR-data_v1_Jun23.xlsx file has been added to the files in the dataset

DATA & FILE OVERVIEW

File List:

PLOS-Dataset_v3_Jun23.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_v3_Jun23.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_v3_Jun23.pdf file.

OSI-Summary-statistics_v3_Jun23.xlsx

This file contains the summary data (PLOS and comparator), some of which is used within the related blog post

<https://theplosblog.plos.org/2023/06/Open-Science-Indicators-Update-Q1-2023>. 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_v3_Jun23.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_v3_Jun23.csv and Comparator-Dataset_v3_Jun23.csv.

OSI-Repository-List_v1_Dec22.xlsx

List of repositories and their characteristics used to identify specific repositories in the PLOS-Dataset_v3_Jun23.csv and Comparator-Dataset_v3_Jun23.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.

OSI-Dimensions-FoR-data_v1_Jun23.xlsx

The data in this file was obtained on 3 May 2023, from Digital Science's Dimensions platform, available at <https://app.dimensions.ai>.

The file lists all the DOIs from both the PLOS-Dataset_v3_Jun23.csv and Comparator-Dataset_v3_Jun23.csv and their corresponding topic classifications in the [Australian and New Zealand Standard Research Classification](#) Fields of Research (ANZSRC FoR) as [implemented by Dimensions](#). This data has been used to carry out the topic-based analysis presented in OSI-Summary-statistics_v3_Jun23.xlsx. The topics were matched to the articles in the PLOS-Dataset_v3_Jun23.csv and Comparator-Dataset_v3_Jun23.csv using the DOI. Topic-based analysis was conducted using the high-level topics denoted by a 2 digit prefix. Articles can have more than one topic assigned.

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_v3_Jun23.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_v3_Jun23.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. Beruria Novich of PLOS for data processing.

DATA-SPECIFIC INFORMATION FOR: PLOS-Dataset_v3_Jun23.csv

Number of variables: 25

Number of cases/rows: 74130

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_v3_Jun23.csv

Number of variables: 25

Number of cases/rows: 8176

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