README for PLOS Open Science Indicators Dataset, version 2

This readme file was generated on 27-3-2023 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 2). 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: 2

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).

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)

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_v2_Mar23.pdf file.

UPDATES ON PREVIOUS VERSION OF THE DATASET

- An additional 9791 articles have been added to PLOS-Dataset_v2_Mar23.csv and 1047 to Comparator-Dataset_v2_Mar23.csv compared to v1.
- Missing information has been standardised as "N/A", replacing some instances of "NA" in the v1 data.
- Dates have been changed from dd/mm/yyyy format in v1 to separate columns for day, month and year in v2.
- Preprint_Match is "TRUE"/"FALSE" in v2 compared to "Yes"/"No" in v1.
- OSI-Country-Region-List_v1_Mar23.csv has been added to the files in the dataset.
- OSI-Methods-Statement_v2_Mar23.pdf has been updated to detail the parameters for the additional articles. More detailed accuracy information has been added.
- OSI-Column-Descriptions_v2_Mar23.pdf has been updated to reflect the changes in date columns.

DATA & FILE OVERVIEW

File List:

PLOS-Dataset_v2_Mar23.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_v2_Mar23.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_v2_Mar23.pdf file.

OSI-Summary-statistics_v2_Mar23.xlsx
This file contains the summary data (PLOS and comparator), some of which is used within the related blog post https://theplosblog.plos.org/2023/04/open-science-indicators/. 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_v2_Mar23.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_v2_Mar23.csv and Comparator-Dataset_v2_Mar23.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.

OSI-Country-Region-List_v1_Mar23.csv
Lists the data from the “Country” columns of PLOS-Dataset_v2_Mar23.csv, the corrections made to these entries and the world region each country maps to. Corrections are mainly due to misspellings and standardising country names (e.g. UK and United Kingdom have been standardised to all be United Kingdom). Some assumptions have been made where the “Country” column contains state or city information instead of country (e.g. “Iowa” has been corrected to United States of America”). Where the country could not be discerned this has been marked as N/A.
The regions used are Africa, Americas, Asia, Australasia, Europe, MENA (Middle East and North Africa).

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_v2_Mar23.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_v2_Mar23.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_v2_Mar23.csv

Number of variables: 25

Number of cases/rows: 71109

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_v2_Mar23.csv

Number of variables: 25

Number of cases/rows: 7635

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