Open Science Indicators Methods documentation for v1 Public Data

This file was prepared on 7-12-2022 by Allegra Pearce and is part of the 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

Assembly of PLOS-Dataset_v1_Dec22.csv:

The entire PLOS Collection was downloaded using the ‘all of PLOS’ API (https://github.com/PLOS/allofplos). We selected a set of 61,318 articles total for the dataset. We initially included articles designated as research articles (article type was “Research Article”, Meta-Research Article”, or Pre-Registered Research Article”), with a publication date between 01/012019 and 30/06/2022. In addition to these criteria we only included articles with a Data Availability Statement identified within the XML file, and at least one of the following sections in the XML: materials|method, and supplementary material. Inclusion of articles with all three section tags was prioritized (i.e. Data Availability Statement, materials|method, and supplementary material). However, approximately 10,000 articles within the final set are missing one of the non-mandatory text section tags (i.e. missing materials|method or supplementary material) and included a full-text analysis to ensure any information provided in an unlabeled section was included in the analysis.

Assembly of Comparator-Dataset_v1_Dec22.csv:

A comparator set of 6588 Open Access articles published in non-PLOS journals was assembled. To ensure a broad subject area match between the PLOS dataset and the comparators, we downloaded the major MeSH terms from PubMed Central (PMC) for the 61,318 PLOS articles. We obtained a list of 11,728 major MeSH terms that appear between 1 and 1083 times in the corpus. Terms that appear on many PLOS articles (e.g. COVID19) correspondingly appear many times in this list. We then randomly selected a 6600-term subset with replacement, such that selected terms can appear multiple times in the created list if they appear frequently in the MeSH distribution.

Articles were chosen for the comparator dataset as follows. We searched within PubMed Central's Open Access corpus (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/) for each term in the 6600-term subset. For each term we excluded articles where the journal title contained the word 'PLOS', and constrained the results to articles of type "Journal Article" published between 2019/01/01 and 2022/06/30; articles already in the comparator dataset were also excluded. A random article was chosen per query term and added to the comparator dataset.

The comparator dataset was then processed with methodology analogous to the PLOS dataset detailed above using the nxml files downloaded from PMC. Due to differences in the provided metadata between PMC nxml and PLOS XML, the metadata collection methods differ between the two corpora. Due to a lack of consistent availability of certain metadata in the PMC nxml files, not all metadata fields were provided per article. An additional field is included in the comparator set to provide further context when interpreting the results: in place of listed disciplines found in the PLOS XML, the list of assigned major MeSH terms is included for each article.

Preprint Detection:

We searched the Crossref database via the Crossref API [https://api.crossref.org/works] for the DOI of each published article. Metadata on article title and the author list was extracted from the Crossref record and used to formulate a search query to find potential preprint records [e.g. bibliographic = article_title, author = article_authors, type = posted-content]. To ensure coverage of articles posted to arXiv, we also searched the DataCite API [https://api.datacite.org/dois] using the same title and author list metadata with the following minor changes: 1) arXiv preprints are not stored under the preprint resource type and therefore no type level filter could be completed, 2) to compensate for querying with no other filters we applied the publisher filter to only include arXiv entries, and 3) due to the strict string match only the family name of each author was used in the query [e.g. titles.title: article_title AND creators.familyName: article_authors AND publisher:"arXiv"].

For each article, the list of potential preprints returned by Crossref was then sorted by the Crossref 'relevance' score (which is a measure of how relevant the preprint is to the search query). Preprint records are classified as 'posted content' in the Crossref API, a category that includes other types of media associated with publications (e.g. published protocols and conference materials). Preprints, as an earlier version of a publication, may have changes to the title or author list than a more recently published protocol (or other content) would not; this may result in a preprint not being the top match when considering all materials. To try to limit matches to non-preprint records we removed records with DOI prefixes that belonged to two organizations that publish other types of content (i.e. protocols.io and Morressier) before evaluation. The author and title, and ORCID ID metadata of the top 20 most relevant results for each article were then used to compute similarity to the published article. The DataCite match process is similar to the Crossref process, with minor differences related to metadata structure and availability: 1) Matching based on ORCID is not possible, as this field is not included in preprint records, and 2) preprint date is recorded as year only for most records.

Title similarity was determined by the Jaccard distance of tokenized titles, if this value was above 0.80 the record was determined to be a match. If the title similarity was greater than 0.10 and the first author's name or orcid ID matched, the article was determined to be a match (see also Cabanac et al. 2019). Potential matches were prioritized by initial search relevance, and the most relevant (i.e. the highest search result to match) record was determined to be the most likely preprint match. For matched preprints we recorded the date of DOI registration, title, author list, as well as the server name and preprint URL (if available). If the server name was not provided the server was estimated from the DOI prefix in the preprint record. If no articles had a similarity above the threshold on either Crossref or DataCite, the article was assigned as having no preprint.

Data and Code Generation:

We first determined if each article had generated one or more datasets to allow consideration of OSIs as both a percentage of all articles as well as for only articles that had shareable datasets, as desired. To do this, we applied a custom Natural Language Processing (NLP) model (https://github.com/DataSeer/dataseer-ml) to the Methods section of the article to detect sentences describing data collection. When the article did not have a detectable Methods section, the full text of the article was analyzed. The model also detects sentences describing the re-use of existing datasets. Since re-analysis of existing datasets frequently requires additional manipulation of the data – and hence the creation of a new shareable dataset – we counted re-use of existing data as 'data generation'.

We detected the generation of shareable code objects with a similar protocol. Sentences in the Methods text of each article were processed by a NLP model designed to detect keywords associated with code generation or script use (e.g. 'script'). An article was also designated as 'generating code' if it mentioned command line software (e.g. Mathematica) or commonly used coding environments (e.g. R or Python).

Data and Code Sharing:

We then assessed whether data were shared within the supplementary files of the article or on an online repository. To determine whether datasets were shared as supplementary files we first excluded image files, specifically files with the mime_type=image or the type .jpg, .tif, .png. We then determined if the file contained data by applying a NLP model to the caption, title, and file type. In addition to this, we used a similar NLP model to analyze sentences from the text in sections where data sharing is usually described (ie. Methods, and Data Availability Statements) to determine if an article shared data on a repository.

We applied a similar workflow to determine whether articles shared code, either as supplemental material or on a public repository. To complement this assessment we also provide DOIs and URLs mentioned in text that are likely to be involved with the code or data sharing. These are taken from text sections that describe sharing and are provided as a complete list of resources shared in the article. We identify commonly used repositories where possible from these URLs and DOIs (see OSI-Repository-List_v1_Dec22.xlsx). We used domain knowledge and frequency of URL domain to identify commonly used online resources; we then verified repositories that hosted code and data before adding them to the detected repository list. This list is not a complete record of every repository used in this dataset, and will continue to be built upon with future data releases.

Accuracy rates

We have aimed for a minimum accuracy rate of at least 85% for all indicators and content sources. The accuracy rate is calculated by randomly selecting 100-200 articles from each corpus and checking them by hand to identify false positives and false negatives. These measures are then used to calculate the overall accuracy of the Dataseer assignments. For PLOS articles, all indicators meet our goal accuracy level but for the comparator corpus data sharing accuracy rates are below this minimum.

Indicator accuracy rates reported by DataSeer.

Indicator	**Accuracy assessment PLOS articles**	**Accuracy assessment Non-PLOS articles**
Data generation	88%	89%
Data sharing	85%	81%
Code generation	85%	92%
Code sharing	97%	94%
Preprint sharing	94%	96%

References:

Cabanac, G., Oikonomidi, T. & Boutron, I. Day-to-day discovery of preprint–publication links. Scientometrics 126, 5285–5304 (2021). https://doi.org/10.1007/s11192-021-03900-7