Open Science Indicators Methods documentation for v7 Public Data

This file was initially prepared on 7-12-2022 by Allegra Pearce and Lauren Cadwallader (version 1), was revised on 06-12-2023 by Lauren Cadwallader and Tim Vines (v5) and updated on 22-05-2024 by Lauren Cadwallader and Scott Kerr (v7). It is part of the dataset:
Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 7). 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_v7_Jun24.csv:

The version 6 PLOS dataset (PLOS-Dataset_v6_Mar24.csv) was combined with new data gathered following the same methodology as version 1. That is, the entire PLOS Collection was downloaded using the 'all of PLOS' API (https://github.com/PLOS/allofplos). We selected a set of 4,510 additional articles with a publication date between 01/1/2024 and 31/3/2024 for the v7 dataset. We initially included articles designated as research articles (article type was "Research Article", "Meta-Research Article", or "Pre-Registered Research Article"). 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, articles missing one of the non-mandatory text section tags (i.e. missing materials|method or supplementary material) are included using a full-text analysis to ensure any information provided in an unlabeled section was included in the analysis.
The v7 dataset has a total of 112,229 articles. All articles were reanalysed using the same algorithm as used in v5.

Assembly of Comparator-Dataset_v7_Jun24.csv:

The version 6 Comparator dataset (Comparator-Dataset_v6_Mar24.csv) was combined with new data gathered following the same methodology as version 1. An additional comparator set of 902 Open Access articles published in non-PLOS journals between 01/1/24 and 31/3/2024 was assembled for v7. The selection method used for versions 1 to 6 was as follows: 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 (v1 dataset). 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 1200-term subset with replacement, such that selected terms can appear multiple times in the created list if they appear frequently in the MeSH distribution. The same list of MeSH terms was used to sample the additional comparator articles for v6. For the new articles included in v7 - articles published between 1/1/24 and 31/3/24, a new set of MeSH terms was used. These were obtained by downloading the major MeSH terms from PLOS articles published between 1/1/22 and 31/12/23. These were used to sample the PMC articles for the comparator set. From this we obtained a list of 10,351 major MeSH terms.

Articles were chosen for the comparator datasets as follows. We searched within PubMed Central's Open Access corpus (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/) for each MeSH term. For each candidate article matching the term, we excluded articles whose XML indicated that the publisher was PLOS., and constrained the results to articles of type "Journal Article" published in the same time period; 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 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.

For the v7 data two changes to the algorithm were implemented and retrospectively run across the entire corpus:

1) the exclusion of protocols.io as a preprint source, as well as other sites that are not preprint repositories. We aimed to exclude protocols.io, American Diabetes Association, Open Access Te Herenga Waka-Victoria University of Wellington, Moressier, Thesis Commons, TMU and Geoscientific Model Development Discussions. If the preprint DOI was for a disallowed server, then we reran preprint gathering. This could result in either no preprint or a different preprint (from an allowed server). For the historical runs, we remediated values as follows: in the case where the preprint DOI contradicted the server (whether or not that server was allowed), we re-looked-up the server using an expanded set of regular expression statements. The same expanded regular expression statements were used to determine if the DOI mapped to the server correctly or not.

2) the removal of postprints (i.e. “preprints” which were published on the same day or after the article was published). In cases where a postprint was eliminated we reran preprint gathering. As a result, some rows are now “NA” while others resulted in a true 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. A more inclusive assessment of data sharing was captured in the “Data_location” column, which assigns data as being shared online, in supplementary material or both. The “online” category includes repositories as well as other

online locations, such as lab websites. It, therefore, includes a greater number of articles although the majority of those sharing “online” are doing so via a repository.

Open Science Indicators Accuracy rates for v5 release

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%

Open Science Indicators accuracy rates for v1 release

Allegra Pearce (updated 10-01-2023)
The accuracy rates presented are for the v1 dataset: Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 1). https://doi.org/10.6084/m9.figshare.21687686.

Below are the calculated accuracy results for the DataSeer analysis and ODDPub (Riedel et al., 2020) for both data and code. We plan to share detailed accuracy results for other indicators, e.g. preprints, in the future. For both the PLOS and Comparator corpus, results are calculated for a 100 article ground truth set manually curated by DataSeer. The manual coding for the accuracy estimates is based on a full human read-through of the article plus testing of

the web links. Data Generation is determined by the presence of one or more data related sentences, either for the generation of new data or the re-use of existing datasets. In each set we've provided accuracy rates, sensitivity, specificity, precision, and F-scores. In addition to this we have provided confusion matrices with the true and false positive and negative labels for each metric (per dataset), these values are what the accuracy measures are based on. Below is a brief definition of each of the accuracy measures.

Accuracy rate (%): proportion of correctly labeled articles

Recall/Sensitivity: ratio of correctly labeled positive cases to total true positive cases

Specificity: ratio of correctly labeled negative cases to total true negative cases

Precision: ratio of correctly labeled positive cases to all cases labeled positive

F-score: harmonized mean of precision and recall (also called sensitivity)

Each of these specialized metrics shows a particular piece of information and is very helpful in diagnosing and directing continual improvements in development.

ODDPub's published F-scores are 0.73 for open data and 0.64 for open code. As a note, the authors also indicate in their publication that the open code assessment (F-score) is likely inaccurate due to the very low occurrence rates of code sharing (11 out of 792, Riedel et al., 2020). The effects of low occurrence rates are also apparent in the PLOS and PMC Comparator corpora studied here

Open science indicators with unbalanced cases (i.e. have many more positive or negative cases than the opposite) can show different impacts per correct or incorrect label in each accuracy metric. Metrics like sensitivity and specificity are a proportion and are sensitive to the total number of true cases. A single incorrect label can have a much larger impact on a proportion when there are fewer cases than when there are many, and as a result a single incorrect/correct label can have a much larger impact on an accuracy metric, while having a much smaller impact on overall accuracy in unbalanced datasets where there are fewer total true cases.

As an example, in data generation (PLOS) there are a total of 11 negative cases with 5 incorrectly assigned as positive cases, as a result the specificity (the ratio of labeled negative cases to total negative cases) is low though the overall impact to accuracy is much smaller (accuracy = 89%, F-score = 0.94). Another example occurs in code sharing (Comparator), where there are 6 true positive cases, with 5 of these correctly labeled as positive, and 5 others incorrectly labeled as positive cases (false positives). The sensitivity is relatively high (0.83), as the majority of the correct cases were labeled correctly, however the precision (and therefore the F-Score which is dependent on the precision and recall/sensitivity) is low (0.50) due to the five false positives. The overall accuracy is still very high in this open science indicator (94%), indicating only a few cases had a strong impact on the precision and F-Score metrics (precision = 0.50, F-Score = 0.64).

These accuracy metrics are excellent tools to give greater context of the strengths and weaknesses of an individual process, but need to be viewed with additional context to gauge the

reliability of the metric. Due to this we prefer to provide accuracy in general which is easier to interpret and is more robust to unbalanced datasets. To give additional context to these metrics we also provide the confusion matrices that have the total of true positive, true negative, false positive, and false negative cases within each set and metric.

PLOS Corpus:

Table 1: Proportion of articles sharing data or code (either in an online repository or as supplemental material), as a proportion of either the number of articles *generating* data or code (Manual Annotation and DataSeer only) or the total number of articles. ODDPub does not estimate whether an article generates data or code, only shares, and so is only included in the second proportion (i.e. sharing/total). These proportions are estimated with the groundtruth subset of the PLOS corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer	ODDPub
Data	68/86 = 79.1%	60/95 = 63.2%	NA
Code	18/41 = 43.9%	16/41 = 39.0%	NA
Sharing/Total	Manual Annotation	DataSeer	ODDPub
Data	68/97 = 70.1%	60/97 = 61.9%	52/97 = 53.6%
Code	18/97 = 18.6%	16/97 = 16.5%	11/97 = 11.3%

Table 2: Accuracy metrics for the PLOS corpus. Results for DataSeer analysis and ODDPub, where applicable, are provided.

DataSeer	Accuracy	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	89%	0.94	0.89	0.99	0.09
Data Sharing	89%	0.92	0.98	0.87	0.96
Code Generation	85%	0.85	0.85	0.85	0.84
Code Sharing	97%	0.83	0.88	0.78	0.97
ODDPub	Accuracy	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Sharing	71%	0.77	0.90	0.67	0.81
Code Sharing	91%	0.69	0.91	0.56	0.99

Table 3: Confusion Matrix of DataSeer and ODDPub detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the groundtruth set of the PLOS corpus manually annotated by DataSeer. ODDPub only evaluates sharing and therefore only has values for data sharing and code sharing. In code sharing totals are displayed removing articles when an annotator is unable to determine if code was used (N = 17).

		DataSeer		ODDpub	
Data Generation		yes	no	yes	no
Manual Annotation	yes	85	1	NA	NA
	no	10	1	NA	NA
Data Sharing		yes	no	yes	no
Manual Annotation	yes	61	9	47	23
	no	1	26	5	22
Code Generation		yes	no	yes	no
Manual Annotation	yes	35	6	NA	NA
	no	6	31	NA	NA
Code Sharing		yes	no	yes	no
Manual Annotation	yes	14	4	10	8
	no	2	77	1	78

Comparator Corpus:

Table 4: Proportion of articles sharing data or code (either in an online repository or as supplemental material), as a proportion of either the number of articles *generating* data or code (Manual Annotation and DataSeer only) or the total number of articles. ODDPub does not estimate whether an article generates data or code, only shares, and so is only included in the second proportion (i.e. sharing/total). These proportions are estimated with the groundtruth subset of the Comparator corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer	ODDPub
Data	45/88 = 51.1%	44/92 = 47.8%	NA
Code	6/44 = 13.6%	10/54 = 18.5%	NA
Sharing/Total	Manual Annotation	DataSeer	ODDPub
Data	45/99 = 45.5%	44/99 = 44.4%	15/99 = 15.2%
Code	6/99 = 6.1%	10/99 = 10.1%	6/99 = 6.1%

Table 5: Accuracy metrics for the Comparator corpus (~6,600 articles). Results for DataSeer analysis and ODDPub, where applicable, are provided.

DataSeer	Accuracy (%)	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	89%	0.94	0.92	0.96	0.3
Data Sharing	81%	0.79	0.80	0.78	0.83
Code Generation	92%	0.94	0.85	1.0	0.85
Code Sharing	94%	0.63	0.5	0.83	0.95
ODDPub	Accuracy (%)	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Sharing	65%	0.43	0.87	0.29	0.96
Code Sharing	98%	0.83	0.83	0.83	0.99

Table 6: Confusion Matrix of DataSeer and ODDPub detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the groundtruth set of the Comparator corpus manually annotated by DataSeer. ODDPub only evaluates sharing and therefore only has values for data sharing and code sharing.

		DataSeer		ODDpub	
Data Generation		yes	no	yes	no
Manual Annotation	yes	85	4	NA	NA
	no	7	3	NA	NA
Data Sharing		yes	no	yes	no
Manual Annotation	yes	35	10	13	32
	no	9	45	2	52
Code Generation		yes	no	yes	no
Manual Annotation	yes	44	0	NA	NA
	no	8	47	NA	NA
Code Share		yes	no	yes	no
Manual Annotation	yes	5	1	5	1
	no	5	88	1	92

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

Riedel, N., Kip, M. and Bobrov, E., 2020. ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications. Data Science Journal, 19(1), p.42. DOI: http://doi.org/10.5334/dsj-2020-042