

Open Science Indicators Methods documentation for v5 Public Data

This file was initially prepared on 7-12-2022 by Allegra Pearce and Lauren Cadwallader (version 1). This revision was prepared 27-11-2023 by Lauren Cadwallader, Tim Vines, Asura Enkhbayar, and Scott Kerr.

It is part of the dataset:

Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 5). <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

The OSI Corpus

The Open Science Indicators (OSI) corpus consists of two separate datasets: articles from all [PLOS Collections](#) and a comparator dataset sampled from the [Pubmed Central Open Access Subset](#) based on the characteristics of the PLOS dataset. The OSI corpus covers articles starting January 1, 2018 and is updated on a quarterly basis.

Table 1. Descriptive statistics for each OSI release

Release	Quarter	PLOS	Comparator	OSI Corpus
v1	2019 Q1 - 2022 Q2	59,708	12,519	72,227
v2	2022 Q3	5,100	1,048	6,148
v3	2022 Q4	4,689	826	5,515
v4	2023 Q1	4,608	967	5,575
v4	2023 Q2	4,232	851	5,083
v5	2023 Q3 plus 2018 Q1-Q4	4,285 + 20,540	857 + 4,110	29,792

In the following sections, we describe the specific article inclusion criteria for both PLOS and comparator dataset, the basic data processing applied to each article independent of the data, code, or preprint indicators, additional metadata retrieved for each article, and a final schema for the OSI corpus.

Updates for v5

An algorithm change implemented in late 2022 affected the estimation of code generation rates, such that about 15% fewer articles were assigned as sharing code from Q3 2022 onwards.

In particular, the original algorithm applied from Q1 2019 to Q2 2022 scored articles that mentioned the term ‘model’ in the Methods section as ‘Generated Code’. Detailed manual analysis of the ground truth article set revealed that ‘model’ was also used in articles that did not generate code, for example in the phrase ‘animal model’. Excluding this term reduced the proportion of articles scored as ‘Generated Code’ by about 15%. Minor adjustments have also been made to the algorithms that detect other OSIs related to data and preprints and these have resulted in minimal changes to the OSI rates.

Collecting the Open Science Indicators has also become more automated between v1 and v5. Manual steps have been eliminated, and the process moved onto an Amazon Web Services cloud server.

This release version also includes data for both PLOS and PMC comparator articles from 2018.

To ensure that data from all quarters are comparable, v5 applies the same algorithm to all quarters, from Q1 2018 to Q3 2023.

Some column headings in both dataset files (PLOS-Dataset_v5_Dec23.csv and Comparator-Dataset_v5_Dec23.csv) have been changed compared to v4 to standardise naming. The Preprint_Match column in both files has been changed to “Yes” or “No” values. Two new columns have been added - “Data_DOIs” and “Quarter”.

Articles

PLOS dataset

The corpus was created in a two-step process. First, the PLOS dataset was created based on a set of inclusion criteria for articles that have been published in each quarter. Then, the comparator dataset was sampled from Pubmed Central (PMC) based on disciplinary characteristics of the PLOS Q1 2019 - Q2 2022 corpus.

The metadata required for these assessments were extracted from the XML versions of the full texts. For PLOS, these files were retrieved using the [allofplos](#) tool while PMC articles were retrieved from the [PMC OA Subset FTP service](#). To assess a PLOS article we applied three criteria which require data extraction from the XML files retrieved using the:

1. **Publication date:** We first determined whether an article was published in the quarter of interest. To do so, we use the electronic publication date which can be found in the XML of an article.
2. **Article type:** Three valid article types were determined: “Research Article”, “Meta-Research Article”, or “Pre-Registered Research Article”. These article type fields are provided in the JATS subject grouping heading attribute¹ for each article.
3. **Article content:** Furthermore, we required articles to contain certain sections in the fulltext to be considered for further processing in OSI. Each article is required to have a *Data Availability Statement* or at least one of the following two sections: *materials/methods* or *supplementary material*. Once again, we processed the XMLs to retrieve and assess this information for each article published by PLOS.

Comparator dataset

In the v4 Comparator dataset, the number of articles was expanded to double the sample in all years from 2019 to the end of Q2 2023. An additional comparator set of 7,847 Open Access articles published in non-PLOS journals was therefore assembled for v4. The selection method used was the same as for the v1 dataset, as described below.

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 appeared on many PLOS articles (e.g. COVID19) correspondingly appeared many times in this list. We then randomly selected a 1200-term subset with replacement, such that selected terms appeared multiple times in the created list if they appeared frequently among the MeSH term list. The same list of MeSH terms was used to sample the additional comparator articles for v4; these same articles were reprocessed for v5.

¹ <https://jats.nlm.nih.gov/publishing/tag-library/1.2/element/subj-group.html>

Using that list of MeSH terms we queried the PMC Open Access subset for candidate articles which were then filtered by a set of inclusion and exclusion criteria until we reach the desired target sample size of 20% of the PLOS dataset. These criteria are:

1. **Publication date:** When retrieving articles from PMC we limit results to the target quarter.
2. **Journal:** Articles published in any PLOS journal were excluded from the comparator dataset.
3. **Article type:** Similar to the PLOS dataset, we retrieved article types from the subject grouping headings. In contrast to the PLOS dataset, these subject grouping headings come from a wide range of different journals and publishers. Therefore, we used heuristics to limit the included articles to original research articles and exclude other scholarly publications such as editorials, retractions, and other journal front matter.
4. **Article content:** As for the PLOS dataset, the same inclusion criteria applied to sections (data availability statements are required and either a methods section or supplementary materials must be present).

Methods

Preprint Detection

We searched the [Crossref](#) database via the [Crossref API](#) 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](#) 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.

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).

An algorithm change implemented in late 2022 affected the estimation of code generation rates, such that about 15% fewer articles were assigned as sharing code from Q3 2022 onwards. In particular, the original algorithm applied from Q1 2019 to Q2 2022 scored articles that mentioned the term ‘model’ in the Methods section as ‘Generated Code’.

Detailed manual analysis of the ground truth article set revealed that ‘model’ was also used in articles that did not generate code, for example in the phrase ‘animal model’. Excluding this term reduced the proportion of articles scored as ‘Generated Code’ by about 15%.

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. Accession numbers are derived by applying a series of regular expression matches to the Data Availability Statement.

Evaluation

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 for DataSeer.

Indicator	Accuracy assessment PLOS articles	Accuracy assessment Non-PLOS articles
Data generation	95%	100%
Data sharing	85%	80%
Code generation	86%	88%
Code sharing	93%	93%
Preprint sharing	94%	96%

Accuracy rates for v5 release

Below are the calculated accuracy results for the DataSeer analysis and ODDPub (Riedel et al., 2020) for both data and code. For both the PLOS and Comparator corpus, results are calculated for a 200 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.

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.

For v5 we have added an additional 100 manually assessed articles to both the PLOS and Comparator Corpus to ensure our accuracy rates remain within our determined threshold and have included the updated tables below.

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 2019 Q1 - 2022 Q2 groundtruth subset of the PLOS corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer	ODDPub
Data	122/193 = 63.2%	139/186 = 74.7%	NA
Code	29/82 = 35.4%	33/76 = 43.4%	NA
Sharing/Total	Manual Annotation	DataSeer	ODDPub
Data	122/194 = 62.9%	139/194 = 71.6%	52/97 = 53.6%
Code	29/194 = 18.6%	33/194 = 17.0%	11/97 = 11.3%

Table 2: Accuracy metrics for the PLOS ground truth corpus.

DataSeer	Accuracy	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	95%	0.98	0.99	0.96	0.00
Data Sharing	85%	0.89	0.83	0.95	0.68
Code Generation	86%	0.82	0.86	0.79	0.90
Code Sharing	93%	0.77	0.73	0.83	0.95
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 (200 articles) and ODDPub (100 articles) detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the 2019 Q1 - 2022 Q2 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	185	8	NA	NA
	no	1	0	NA	NA
Data Sharing		yes	no	yes	no
Manual Annotation	yes	116	6	47	23
	no	23	49	5	22
Code Generation		yes	no	yes	no
Manual Annotation	yes	65	17	NA	NA
	no	11	101	NA	NA
Code Sharing		yes	no	yes	no
Manual Annotation	yes	24	5	10	8
	no	9	156	1	78

PMC 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 2019 Q1 - 2022 Q2 groundtruth subset of the PMC Comparator corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer	ODDPub
Data	77/200 = 38.5%	79/201 = 39.3%	NA
Code	16/77 = 20.8%	16/84 = 19.0%	NA
Sharing/Total	Manual Annotation	DataSeer	ODDPub
Data	77/201 = 38.3%	79/201 = 39.3%	15/99 = 15.2%
Code	16/201 = 7.9%	16/201 = 7.9%	6/99 = 6.1%

Table 5: Accuracy metrics for the 2019 Q1 - 2022 Q2 PMC Comparator corpus. Results for DataSeer (200 ground truth articles) analysis and ODDPub (100 ground truth articles), where applicable, are provided.

DataSeer	Accuracy (%)	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	100%	1.00	1.00	1.00	0.00
Data Sharing	80%	0.74	0.73	0.75	0.83
Code Generation	88%	0.84	0.81	0.88	0.87
Code Sharing	93%	0.56	0.56	0.56	0.96
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 (200 ground truth articles) and ODDPub (100 ground truth articles) detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the 2019 Q1 - 2022 Q2 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	200	0	NA	NA
	no	1	0	NA	NA
Data Sharing		yes	no	yes	no
Manual Annotation	yes	58	19	13	32
	no	21	103	2	52
Code Generation		yes	no	yes	no
Manual Annotation	yes	68	9	NA	NA
	no	16	108	NA	NA
Code Share		yes	no	yes	no
Manual Annotation	yes	9	7	5	1
	no	7	178	1	92

Q2 2023 Accuracy Measurements

To ensure that the algorithm performs well with more recent articles, we created a second ground truth article set for 100 articles from Q2 2023. The results tables are given in the same order as above.

PLOS corpus

Table 7: 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. These proportions are estimated with the groundtruth subset of the Q2 2023 PLOS corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer
Data	71/101 = 70.3%	78/101 = 77.2%
Code	20/43 = 46.5%	20/33 = 60.6%
Sharing/Total	Manual Annotation	DataSeer
Data	71/103 = 68.9%	78/103 = 75.7%
Code	20/103 = 19.4%	20/103 = 19.4%

Table 8: Accuracy metrics for the PLOS Q2 2023 ground truth corpus.

DataSeer	Accuracy	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	96%	0.98	0.98	0.98	0.00
Data Sharing	89%	0.93	0.88	0.97	0.72
Code Generation	86%	0.82	0.94	0.72	0.97
Code Sharing	84%	0.60	0.60	0.60	0.90

Table 9: Confusion Matrix of DataSeer detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the Q2 2023 ground truth set of the PLOS corpus manually annotated by DataSeer.

		DataSeer	
Data Generation		yes	no
Manual Annotation	yes	99	2
	no	2	0
Data Sharing		yes	no
Manual Annotation	yes	69	2
	no	9	23
Code Generation		yes	no
Manual Annotation	yes	31	12
	no	2	58
Code Sharing		yes	no
Manual Annotation	yes	12	8
	no	8	75

PMC Comparator Corpus

Table 10: 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. These proportions are estimated with the Q2 2023 ground truth subset of the Comparator corpus manually annotated by DataSeer.

Sharing/Generating	Manual Annotation	DataSeer
Data	60/119 = 50.4%	63/120 = 52.5%
Code	18/68 = 26.4%	13/67 = 19.4%
Sharing/Total	Manual Annotation	DataSeer
Data	60/120 = 50.0%	63/120 = 52.5%
Code	18/120 = 15.0%	13/120 = 10.8%

Table 11: Accuracy metrics for the Q2 2023 PMC Comparator corpus.

DataSeer	Accuracy (%)	F-Score	Precision	Recall (Sensitivity)	Specificity
Data Generation	99%	1.00	0.99	1.00	0.00
Data Sharing	81%	0.81	0.79	0.83	0.78
Code Generation	81%	0.83	0.84	0.82	0.79
Code Sharing	89%	0.58	0.69	0.50	0.96

Table 12: Confusion Matrix of DataSeer detection results for generation and sharing of research products (either in an online repository or as supplemental material). Results are shown for the Q2 2023 ground truth set of the PMC Comparator corpus manually annotated by DataSeer.

		DataSeer	
Data Generation		yes	no
Manual Annotation	yes	119	0
	no	1	0
Data Sharing		yes	no
Manual Annotation	yes	50	10
	no	13	47
Code Generation		yes	no
Manual Annotation	yes	56	12
	no	11	41
Code Share		yes	no
Manual Annotation	yes	9	9
	no	4	98

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>