

Open Science Indicators: Protocol Sharing

Technical Documentation

This file was prepared on 27-9-2023 by Asura Enkhbayar and is part of the dataset: Public Library of Science (2022) PLOS Open Science Indicators. Figshare. Dataset (version 4). <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: ihrynaszkiwicz@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

This document provides an overview of the preliminary version of the *Protocol Sharing* indicator and includes documentation of the conceptual model and technical components.

Event/Relationship Model

The underlying model of the *Protocol Sharing* indicator is similar to the [Event Data](#) (or Relationships) model used by Crossref for their Event Data service. The following graph shows the three main components which model the basic triad of protocol sharing which is depicted in the following diagram.

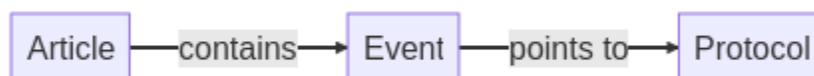


Fig 1. Research articles are linked to protocols through sharing events

Each of the components in the event model are defined as:

- **Research article:** A published research article with common structural elements such as references, links to relevant resources, or supplemental materials

- **Sharing event:** The mentioning of a protocol in a research article in the form of a citation, hyperlink, or supplemental material
- **Protocol:** Detailed and/or step-by-step instructions for carrying out a research procedure, which are findable and accessible on the internet

Technical components and processing pipeline

Each component of the event model relates to a module and processing pipeline. This section provides a more in-depth technical description of these modules. Following the conceptual model outlined above the following section provides the technical implementation and processing steps for each module. Each section is concluded with the respective data schema which, collectively, describe the data model produced by this processing pipeline. It is important to note that the public data releases are specific views of this model with a subset of available columns.

Articles

The minimal input for this module is a list of DOIs representing the input sample. Each input DOI is then enriched with additional data from external sources. For this preliminary data release, all 78,363 research articles published by PLOS from Q1 2019 through Q2 2023 have been sampled. A proportional comparator set of 18,000 PMC articles was sampled using MeSH terms. More details can be found [here](#).

This dataset of articles is then enriched with additional metadata from several scholarly metadata aggregators. Crossref Metadata is used for basic article metadata and reference data. [OpenAlex](#) is currently not used but author disambiguation and concept data are potential future improvements. Finally, the [NCBI ID converter API](#) provides a way to query the most up-to-date PMIDs & PMCIDs for articles which often take a while to propagate through metadata infrastructure.

Article Data Schema

Table 1 describes the data schema for articles in this model with column descriptions and provenance information.

Table 1. Data schema for articles

#	Column	Description	Provenance
1	source	PLOS or PMC	Sampling method
2	doi	Article DOI	Sampling method
3	pmid	Retrieved PMID	NCBI
4	pmcid	Retrieved PMCID	NCBI

5	year	Publication year	Crossref
6	month	Publication month	Crossref
7	day	Publication day	Crossref
8	title	Publication title	Crossref
9	container	Publication source (e.g. journal)	Crossref
10	publisher	Publisher	Crossref
11	article_type	Article type	Crossref
12	subject	Subject	Crossref
13	pub_date	A clean publication date constructed from publication year, month, and day	DataSeer
14	quarter	The publication quarter derived from the pub_date	DataSeer

Protocols

The protocol finder module collects protocols based on an allowlist of 12 sources. This list of sources is non-exhaustive and has been determined in cooperation with PLOS.

Table 2. Allowlist of selected protocol sources

#	Name	ShortName	Type
1	protocols.io	protocolsio	Repository
2	Protocol Exchange	prot_exchange	Repository
3	Bio-protocol	bio_prot	Publication, Repository
4	STAR Protocols	star_prot	Publication
5	MethodsX	methodsx	Publication
6	Journal of Visualized Experiments	jove	Publication
7	BioTechniques	biotechniques	Publication
8	Cold Spring Harbor Protocols	coldspring	Publication
9	Current Protocols	current_prot	Publication
10	Nature Protocols	nature_prot	Publication
11	PLOS ONE	plos	Publication

For each of these sources we collect all available items that Crossref returns in order to compile an initial list of protocol candidates. Not every item that has been published by one of these sources will meet our definition of a protocol/protocol article.

- **Protocol candidate:** A work that is returned by Crossref when queried for one of the selected protocol sources
- **Protocol:** A protocol candidate that also meets additional filtering criteria

Collecting protocol candidates

The 12 protocol sources differ in their nature as some are peer-reviewed journals, collections of edited books, or open repositories without peer-review. Additionally, some of these sources are owned and operated by organizations as their only offered service while others are part of a larger portfolio. Therefore, the collection processes differ across some of the protocol sources:

- **Journals & Repositories.** For most protocol sources (#1 - #10 in the table above) a combination of API queries were used to retrieve associated works, i.e., protocol candidates. E.g., for BioProtocol Journal we used three separate queries to find as many works as possible associated with it: (1) a query using the Crossref member ID 9223, (2) a query for all works with the container_title BIO-PROTOCOL, and (3) one for all works with the group_title BIO-PROTOCOL. Each of these combinations of queries for these sources was curated manually and, if available, the total count of returned works was compared with total protocol counts reported by publishers.
- **Springer Protocols.** The Springer Protocols collection was processed differently as these protocols are book chapters and the previous queries only work for journals and repositories. Using the list of book titles provided by Springer ([link](#)¹) we then queried Crossref for works with the ISSNs provided by Springer.
- **PLOS ONE.** The article type metadata provided by Crossref is not sufficient to determine whether a PLOS publication is a protocol as the article subtype is only available in the web view or fulltext XML. Therefore, the article subtype was extracted from the PLOS fulltexts retrieved via [allofplos](#) to then filter for the Lab Protocol article type.

Filtering protocols

The resulting list of protocol candidates contains a wide range of works including desired protocols but also editorials, corrections, and errata, but also review articles and other academic writing which does not contain detailed and/or step-by-step instructions for carrying out a research procedure.

¹ last accessed 19.07.2023

A first easy filtering step is to remove works with invalid article types in Crossref. Valid types are considered to be `journal-article`, `book-chapter`, `posted-content`, and `book` while others such as `dataset`, `report`, or `component` are excluded.

Where possible, each source is also refined based on DOI patterns and published article types. The following table provides an overview of all applied DOI patterns and their rationales for each protocol source.

Table 3. DOI patterns which were used to filter protocol candidate for each protocol source

#	ShortName	DOI pattern	Rationale
1	protocolsio	10.17504	ZappyLab registers DOIs for protocols as protocols.io; the DOI prefix is enough to identify protocols
2	protocol_exchange	10.21203/rs.3	ResearchSquare's initial registered DOI prefix
3	protocol_exchange	10.21203/rs.2	ResearchSquare's initial registered DOI prefix
4	protocol_exchange	10.1038/protex	DOI prefix after purchase by Springer
5	bio_prot	10.21769/bioprotoc	DOI pattern for peer-reviewed protocols
6	bio_prot	10.21769/bio	DOI pattern for peer-reviewed protocols
7	bio_prot	10.21769/p	DOI pattern for preprints
8	bio_prot	10.21769/l	DOI pattern for preprints
9	star_prot	10.1016/j.xpro	Elsevier prefix and journal identifier
10	methodsx	10.1016/j.mex	Elsevier prefix and journal identifier
11	jove	10.3791	JoVE; the DOI prefix is enough to identify protocols
12	biotechniques	10.2144/btn	Future Science prefix and journal identifier
13	coldspring	10.1101/pdb.prot	CSHL Press prefix, journal identifier, and article type prefix (<code>prot</code>) for protocol articles
14	current_prot	10.1002/cpz1	Wiley prefix and journal identifier
15	nature_prot	10.1038/s41596	Nature prefix and journal identifier
16	nature_prot	10.1038/nprot	Nature prefix and journal identifier
17	nature_prot	10.1038/nport	Nature prefix and journal identifier (it's a

typo)

Springer protocols are found under two separate DOI prefixes (10.1007 and 10.1385) but as they are part of a collection (of books) there is no unique DOI pattern to identify protocols. Instead, we make sure that book chapters are part of one of the identified books by matching DOIs (found in the same spreadsheet published by Springer²).

Lastly, PLOS does not require additional filtering as protocols are sufficiently identified by the article subtype in the previous step, i.e., all PLOS protocol candidates are protocols.

Processing protocols

We use OpenAlex's [locations](#) field to add alternative URLs at which versions of this protocol can be found. E.g., a protocol with the DOI_A is accessible at https://doi.org/DOI_A (URL1) which will typically resolve to a publisher landing page https://publisher.com/PROPR_ID_A (URL2) and additionally it might also have been previously published as a preprint at https://preprint.server.org/DOI_B (URL3). Each of these URLs would be added to the allowlist alongside the original DOI of a protocol.

All DOIs and URLs are also cleaned and normalized to ensure that we can successfully identify matches between sharing events in articles and protocols. DOIs are lowercased and any leading and trailing whitespace characters are removed. URLs are normalized using the [hyperlink](#) package. Additionally, the URL scheme is also removed under the assumption that both [http](#) and [https](#) link to the same resource.

Protocol Data Schema

The following table describes the data schema for protocols in this model including column descriptions and provenance information:

Table 4. Data schema for protocols

#	Column	Description	Provenance
1	doi	Protocol DOI	Crossref
2	source	Name of the protocol source	Crossref
3	source_type	A classification of protocol sources	DataSeer
4	article_type	Article type of the protocol	Crossref
5	title	Title	Crossref
6	publisher	Publisher	Crossref

² <https://www.springernature.com/gp/librarians/products/databases-solutions/springerprotocols>

7	url	The main URL reported by Crossref	Crossref
8	urls	The list of additional URLs (incl. url)	Crossref, OpenAlex
9	urls_norm	The list of normalized URLs	DataSeer
10	urls_stripped	The list of normalized URLs with stripped schemes	DataSeer
11	pub_date	A clean publication date constructed from publication year, month, and day	DataSeer
12	quarter	The publicatoin quarter derived from the pub_date	DataSeer
14	is_oa	Whether the protocol is open access	OpenAlex
15	oa_status	Open Access status (i.e., OA color)	OpenAlex

Events

Having identified the relevant set of research articles and protocols, we now need to establish when and how those articles shared any of the protocols we identified. We distinguish between three fundamental types of sharing events, identified as relevant in previous research by PLOS³:

- **References:** A research article formally cites a protocol. Therefore, this protocol should be found in the references list of the article.
- **Links:** A research article links to a protocol in the main body text of the article. The link might or might not include the DOI of the protocol.
- **Supplemental information:** A research article shares a protocol by appending it as a file in the article's supplemental information (SI).

In order to extract and assess the links and SI files the fulltext of the research article is required. For PLOS, the [allofplos](#) package was used to download available XML versions and for PMC the [eutils](#) suite was used to download the XMLs.

Processing three types of events

References. As we previously downloaded the available metadata from Crossref for every article in the corpus, we only need to extract each reference from the Crossref responses. It is important to note that publishers do not always submit these references with DOIs but Crossref attempts to match references with their respective DOIs. However, some references will remain without a DOI and we are currently NOT matching these items ourselves.

Both PLOS and PMC articles are in the [JATS](#) format which standardizes the fundamental structure of articles and naming conventions for basic elements. Within the PMC corpus the

³ <https://osf.io/preprints/metaarxiv/7jxav>

articles come from a variety of publishers and hence there are differences in both the implementation and version of JATS versions. Keep in mind that the following descriptions of XML processing are simplified for brevity.

Links. For each article we extract all URLs from the body of the text in `<ext-link>` elements. Additionally, if available, we attempt to save section information, i.e. in which section of an article the link was shared. For PLOS articles we also have to discard all links found in the `review-history` section of articles which have agreed to publish the peer-review process as that data is not part of the actual article but can nevertheless contain relevant links.

SI files. Finally, if available, we extract SI files for each article appended in SI with some metadata such as title, caption, or descriptions. For PLOS, the metadata quality is high and consistent but for PMC it is important to remember that SI files often do not contain any metadata. A further challenge is that publishers often do not enforce a consistent use of the caption or description fields. Therefore, SI file metadata is less reliable for the PMC comparator set.

Matching articles, events, and protocols

Every article in the corpus is now associated with a list of sharing events which are either references, links, or SI files. What remains to do is determine which of these sharing events are in turn associated with a protocol in our allowlist of protocols.

References. If the DOI of a reference event matches with one of the DOIs in our allowlist we consider that event to be sharing a protocol.

Links. If the URL found in a link event matches with one of the URLs (retrieved from OpenAlex) of a protocol in our allowlist we consider that event to be sharing a protocol.

Supplemental Information (SI) files. SI events are assessed based on a denylist (see table below) approach with a set of invalid protocol terms curated by PLOS and DataSeer. First, only SI files which contain the term protocol in the first place are considered a candidate event. If the title of a candidate event (or title and caption for PMC) contains any term in the denylist the SI file is removed. The remaining events are considered to be sharing a protocol as part of the Supplemental Information.

Table 5. Denylisted terms for protocols

Denylist category	Excluded terms
Clinical study protocols	study protocol; clinical protocol; trial protocol; research protocol; prospective protocol
Systematic review protocols	review protocol; systematic review protocol; prospero protocol; meta-analysis protocol
Preregistered protocols	preregistration protocol; preregistered protocol; registered protocol

Ethics protocols ethics protocol; irb protocol

Event Data Schema

The following table describes the data schema for sharing events in this model including column descriptions and provenance information:

Table 6. Data schema for sharing events

#	Column	Description	Provenance
1	event_id	An ID for each event	DataSeer
2	cohort	PLoS or PMC	DataSeer
3	source	DOI of the article where the event was found	DataSeer
4	type	Event type (citation, link, SI)	DataSeer
5	target	ID of the target which the event was pointing to	DataSeer, Crossref
6	target_doi	DOI of the target (optional)	DataSeer, Crossref
7	target_url	URL of the target (optional)	DataSeer, OpenAlex
8	location	Name of the protocol source	DataSeer
9	location_type	Type of the protocol source	DataSeer

Performance & limitations

The assessment of the quality of the Protocols indicator proved challenging as the events we are measuring occur less frequently than they do with other OSI indicators like data sharing or code sharing. Therefore, in order to assess the quality of this preliminary release we opted to sample from articles that were identified as sharing protocols in an earlier version of this model. By doing so, we were able to focus on the finer details and diversity of protocol sharing behaviors rather than being concerned with automated detection of shared protocols in the wild.

Disclaimer: The reported statistics and numbers are not representing a real-world performance of this model but need to be understood in the context of a detailed investigation of a set of carefully sampled articles.

200 articles were manually assessed for protocol sharing (yes/no). Initially, we sampled 100 articles with a 50/50 split between articles that did and did not share protocols, as assessed by

a pre-release version of the model, for both PLOS and PMC⁴. These 200 articles evenly split between PLOS and the comparator dataset makes up the ground truth (GT) sample which was annotated by curators.

However, in order to fairly assess the performance of our model we have to consider that our chosen approach of identifying protocols based on allowlisting is not designed to find and identify protocol sharing outside of the scope of the 12 sources. Therefore, the performance assessment is split into two parts:

- **Performance of allowlisting.** How does the model perform if GT is also limited to allowlisted protocols?
- **Overall performance.** How does the model perform if we include all other protocols in GT which the model was not designed to find?

Performance of allowlisting

As the model will always only find protocols in our allowlist, the predicted values do not change for this assessment. The GT data, on the other hand, is limited to protocols found during assessment and were later matched to our allowlist. Therefore, the number of protocol sharing articles in the GT will be lower than the actual number found by our curators.

Of the 46 articles which we predicted to share a protocol, this filtered GT set finds 26 of them (56.5%) and one additional article which our algorithm did not identify. Closer inspection shows that the protocol in question was not found because the reference was missing a DOI and Crossref (and OpenAlex) failed to match it to an existing DOI in the system. This reliance on external data is an inherent limitation but it's reassuring that the occurrence seems to be low. Further options to scope this problem could be to investigate the quality and performance of reference matching by Crossref and OpenAlex.

Table 7. Confusion matrix for protocol sharing predictions for 100 PLOS articles limited to protocols from the allowlist. Our model predicts 46 articles as protocol sharing while the GT assessment only found 27 articles which shared a protocol from the allowlist. F1-score: 0.71, accuracy: 0.79, precision: 0.57, sensitivity: 0.96, specificity: 0.73

		Prediction		
		Yes	No	
Actual	Yes	26	1	27
	No			

⁴ Due to minor updates to the model, in the final GT dataset the current model classifies 46 articles in each PLOS and PMC as protocol sharing. Therefore, in what follows the number of articles sharing protocols will sum up to 46 rather than 50.

PLOS		Prediction		
		Yes	No	
	No	20	53	73
		46	54	100

The largest difference between GT and our predictions are the 20 articles which the curators determined to be not-sharing while our algorithm identified citations or links to protocols from our whitelist. A closer investigation of these 20 articles reveals an important area for future research and work which is the distinction between protocol citation and protocol use.

Events in context: Protocol citation vs. protocol use

20 of 46 (43.4%) articles that we predicted to share protocols are actually not using these protocols in the context of their methodology. Instead, these protocols are cited in introductions and other sections outside of the Methods.

Our model labels [this article](#) as protocol sharing as it references [this Springer protocol](#). However, if we look at the actual in-text mention of the protocol we can see that it wasn't referenced for its methodological contribution:

These sulfated polysaccharides always occur as mixtures in tissues with individual components varying slightly in stereochemistry, length, and sulfation pattern (5).

Another example is [this Nature protocol](#) which is mentioned in the introduction of this [PLOS article](#):

Currently, influenza vaccine production heavily relies on traditional embryonated egg technology [5].

This further demonstrates a limitation of this approach which only considers citations, links, and SI as sharing events without their context in the article (i.e., a connection between DOI/URLs rather than a (con)textual event). This could be tackled in future by classifying the sentence as either one that describes protocol use in the study that the article reports or one that does not (cf the quotes above).

Overall performance

We now move on to the actual GT assessment that was produced by our curators which includes protocols outside of the 12 allowlisted sources. This leads to an increase of protocol sharing articles in GT (69 instead of 27 previously). While it might be tempting to assume that

this should inversely correlate with model performance (more protocol-sharing articles in GT -> more protocol-sharing articles missed) it is important to look at the confusion matrices and accuracy scores to make a comparison.

Table 8. Confusion matrix for protocol sharing predictions for 100 PLOS articles. F1-score: 0.80, accuracy: 0.77, precision: 1.00, sensitivity: 0.67, specificity: 1.00

PLOS		Prediction		
		Yes	No	
Actual	Yes	46	23	69
	No	0	31	31
		46	54	100

The GT assessment showed that 69 of 100 PLOS articles shared protocols. Our model correctly identified 46 of 69 (67%) articles as protocol sharing. It correctly identified all 31 articles which did not share protocols. The model also produced no false positives. However, 23 false negatives, articles which share protocols which we could not find with our model. Therefore, these 23 articles (and the missed protocols) can provide insights into the quality and limits of our allowlist of 12 sources.

Beyond our allowlist: Protocols outside of the 12 sources

We are missing a third of all articles that shared a protocol (23/69). While this could indicate that our allowlist of protocol sources needs to be expanded, a closer inspection of the data shows that this might be an inherent limitation of this approach. Figure 2 shows a graph of all containers (journals, books, platforms) that were shared in the GT sample ordered by the number of protocols.

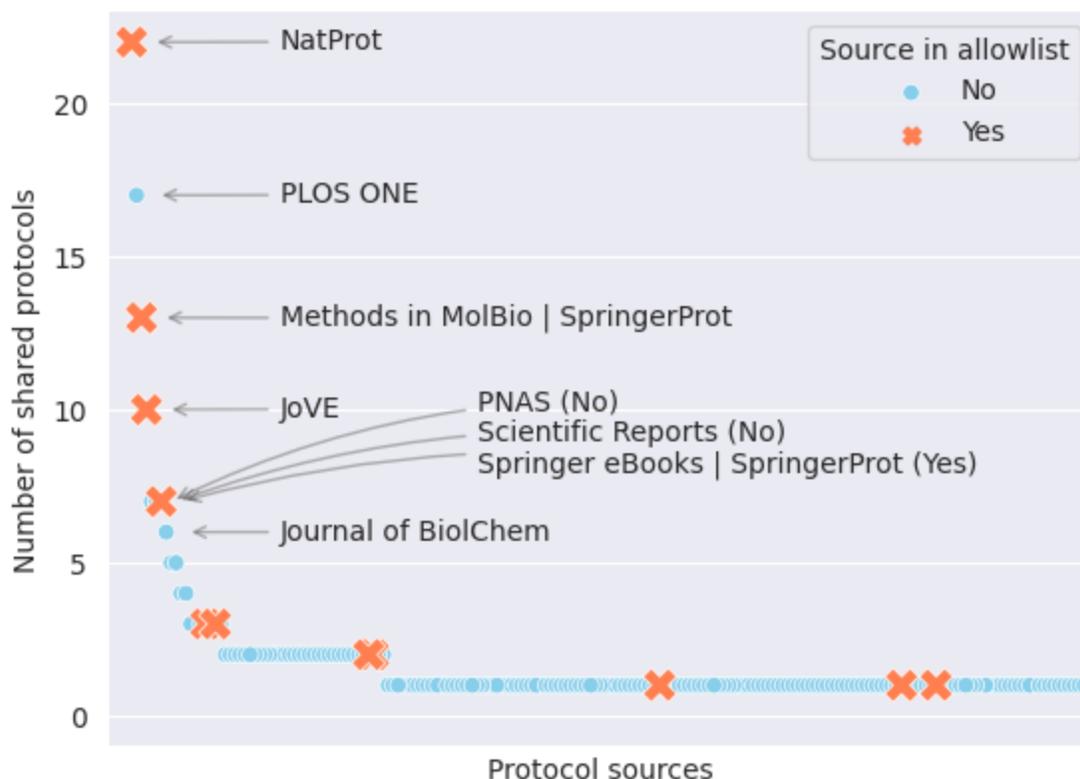


Fig 2. Protocol sources found in GT ordered by their number of shared protocols. Protocol sources that are in our allowlist (Methods in Molecular Biology & Springer eBooks are part of the Springer Protocols collection) are highlighted in orange. The allowlisted PLOS ONE appears as a blue dot as these articles are not Lab Protocols and therefore missed by our protocol collection.

As expected the distribution is highly skewed with 195 journals with an average of 1.7 protocols per source. Our allowlist covers a good amount of sources in the upper part of this list. The remaining, highly-shared sources do not necessarily specialize in publishing methods or protocols articles (e.g., PLOS ONE, PNAS, Scientific Reports, and Journal of Biological Chemistry are among the top 8 sources). Therefore, a location-based approach as ours would not be able to identify these protocols. The use of more consistent article-level metadata for published protocols is one possible solution, but would require community consensus.

Results for the comparator set

The results for the PMC comparator set are very similar to the ones for PLOS reported above. As we've already discussed the most interesting insights for PLOS we won't go into further detail for the comparator set at this point.

- Table 9. (A) Confusion matrix for protocol sharing predictions for 100 PMC articles limited to protocols from the allowlist. F1-score: 0.80, accuracy: 0.77, precision: 0.98, sensitivity: 0.67, specificity: 0.97
 (B) Confusion matrix for protocol sharing predictions for 100 PMC articles. F1-score: 0.80, accuracy: 0.77, precision: 0.98, sensitivity: 0.67, specificity: 0.97

(A) PMC

		Prediction		
		Yes	No	
Actual	Yes	29	1	30
	No	17	53	70
		46	54	100

(B) PMC

		Prediction		
		Yes	No	
Actual	Yes	46	23	69
	No	0	31	31
		46	54	100