Semi-automated extraction of information on open datasets mentioned in articles

Anastasiia Iarkaeva[1], Evgeny Bobrov[1], Jan Taubitz[1], Benjamin Gregory Carlisle[1], Nico Riedel[1]
[1]Berlin Institute of Health at Charité (BIH), QUEST Center for Responsible Research

May 12, 2022
3
dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

evgeny.bobrov

This protocol describes how to determine for a body of research articles, whether underlying datasets have been openly shared. Statements on shared data are detected within articles using the ODDPub text mining algorithm, and are then further processed using an openness extraction form implemented in Numbat. This extraction form was developed to guide and document the manual validation of automatedly detected Open Data statements. For one article, several datasets are checked, one per dataset location. The extraction form consists of checks of data availability and reusability, loosely inspired by the FAIR principles. The resulting table gives an overview of, amongst others, dataset location, applied license, and data formats. Data sharing in supplements, data reuse and restricted data sharing are also documented as alternatives to open data.

DOI
dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel 2022. Semi-automated extraction of information on open datasets mentioned in articles. **protocols.io** https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

open data, screening tools, data reuse, data sharing, semi-automated, FAIR data, open science, ODDPub, Numbat, data availability

protocol ,

Mar 03, 2022

May 12, 2022

59031

- List of articles for which you want to determine the openness of the corresponding datasets
- R studio to run ODDPub (both are open source)
- Numbat software to run the Openness extraction form

Screen a set of publications for statements indicating Open Data

1 Collate a list of article identifiers

Start from a list of articles, for which you want to determine openness of datasets underlying these articles. Those can typically be obtained by **searching publication databases** (e.g. Pubmed, Web of Science, Dimensions, Embase) for the search criteria of interest and exporting the results with the relevant metadata fields.

Search criteria could be institutional affiliations of authors, specific fields of research, specific journals, and typically a publication date range. If the focus is on articles from one institution, it is also possible that there is a curated list of publications provided by the institution itself (often provided by the institutional library).

The development and optimization of ODDPub were geared to Open Data criteria similar to those linked further below. However, some changes in the criteria were introduced over time.

2 Obtain the article full texts

There are multiple options for this:

- via PubMed Central (Open Access articles)
- via the full-text links provided by the unpaywall API (Open Access articles).
- APIs for full-text retrieval of subscription-based articles, offered by several big publishers, e.g. Elsevier, Wiley, or Springer/Nature
- Additionally, the full-text R package offers a solution that combines several of those data sources for downloading full texts.

All those retrieval options will be limited to the articles that can be accessed either as Open Access articles or via the subscriptions provided by the institution at which the article full texts are retrieved.

Store all retrieved article full-texts either as PDFs or as text files (e.g. XML) in one folder.

3 Apply ODDPub to article full texts

ODDPub (Open Data Detection in Publications) is an Open Source text mining algorithm implemented in R. It screens article publications for data sharing statements throughout the article, using keywords and keyword combinations defined in the script. For this, the publications have to be prepared in PDF or text file format and stored in a local folder. Only publications with full-text access will be screened, as these are needed as input to ODDPub. There is no limit to the number of publications.

The workflow of ODDPub includes **three steps**:
1. Generation of **text files** out of PDF files (only if the full text was not already obtained in a text format, e.g. XML).
2. Search for the keywords, defined in the script, such as *repository name*, *data availability statement*, or *reference to supplementary material*. The words could be found both within and across the sentence borders (spread throughout the larger paragraph).
3. Matching of detected keywords and regular expressions (from the script). If any keyword group matches, the publication is categorized as Open Data.

Besides the detection of Open Data statements, the algorithm detects the location of shared data (general-purpose repository, field-specific repository, supplement), as well as statements indicating Open Code (open source software).

Please refer to Riedel et al. (2020) for details on the development, functionality, validation, and performance of ODDPub.

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

3.1 First, install ODDPub in R using the following command:

Installation of ODDPub

```
# install.packages("devtools") # if devtools currently not installed
devtools::install_github("oddpub-folder")
```

Use e.g. RStudio.

You will need a **poppler** library. Find the installation instruction on:
- Windows - https://blog.alivate.com.au/poppler-windows/
- Mac - https://formulae.brew.sh/formula/poppler

As a rule, you do not need the installation of poppler on Linux - check the status on your device.

3.2 Then, set the working directory to the folder with all examined publications. Here we've named it 'PDFs':

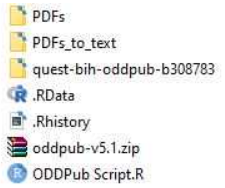

Fig. 3.2. Folder of the ODDPub algorithm on Windows OS

3.3 After you've successfully installed ODDPub and stored the PDFs in one place, run the following command to start evaluation for Openness:

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

ODDPub Script

```
library(tidyverse)
library(oddpub)

oddpub::pdf_convert("PDFs/", "PDFs_to_text/")
PDF_text <- oddpub::pdf_load("PDFs_to_text/")

oddpub_results <- oddpub::open_data_search(PDF_text)
oddpub_results$article <- oddpub_results$article %>%
str_remove(fixed(".txt")) %>%
str_replace_all(fixed("+"), "/")
write_csv(oddpub_results, "oddpub_results.csv")
```

An R-Script to run the Open Data and Open Code detection via the ODDPub algorithm.

After the screening, an **output** in *.csv* format will be created in the same folder as the ODDPub script. You can change the format to any other needed (e.g. .txt, .tsv).

- In the example below, the article with DOI *10.1101/gr.275995.121* (*row #4*) represents a **false-positive result** for **Open Data**, which will be detected during manual validation, as described further below (Sections 4-8).
- The positive result for **Open Code** (row #4), as well as the positive result for **Open Data** for DOI *10.1126/scitranslmed.abe8952* (*row #9*) are **true positives**, which will also need to be confirmed for Open Data by manual validation. Open Code is not pursued further.

A	B	C	D	
article	is_open_data	open_data_category	is_open_code	open_data_statem
10.1186/s13054-022-03894-5	FALSE		FALSE	
10.1002/ana.26326	FALSE		FALSE	
10.1101/gr.275995.121	TRUE	general-purpose repository	TRUE	software availabili variant sets as wel website for the int deletions insertion available at the cadd-sv webserver (https://cadd-sv.bi as well as zenodo (https://doi.org/10
10.1002/jcsm.12927	FALSE		FALSE	
10.1001/jamaneurol.2021.5321	FALSE		FALSE	
10.1016/j.redox.2022.102242	FALSE		FALSE	
10.1073/pnas.2105691119	FALSE		FALSE	
10.1126/scitranslmed.abe8952	TRUE	field-specific repository, supplement	FALSE	rnaseq data have b arrayexpress datal (www.ebi.ac.uk/ar accession number tables s1 to s16 da protocol for this pa

Tab. 3.3. Example result after running ODDPub - two cases of Open Data and one of Open Code were found.

Numbat installation

4 What is Numbat and how to install it

Numbat is used to extract information from primary sources for the purpose of writing systematic reviews in an academic context and manage the resulting databases. This includes assigning extractions to raters and reconciliation of outcomes between raters (i.e., comparison and consolidation into a joint assessment).

A short description of Numbat functionalities:

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

- In principle, you create an extraction form (in other words, a questionnaire) that corresponds to your requirements (in this case a list of questions about datasets to be checked), which will be repeated for each record.
- You then import your detected information (in this case from ODDPub) into Numbat (a .csv table is required) which contains the records you want to validate.
- If you want to compare extraction between raters, you can assign the dataset to several users, otherwise only to one responsible person.
- The resulting dataset with all answers in the extraction form can be exported as a table.

Numbat cannot compute any statistics, it is set up to gather information about records that without using a semi-automated extraction form would have to be entered completelly manually.

It is a PHP-based software that is free and open-source under the GNU AGPL v 3 license.

Here is a short instruction on how to install your own Numbat software (under Windows OS).

You will need:
- Apache HTTP Server
- MySQL Database - provided by your company
- PHP

4.1 Clone the Numbat repository from GitHub locally. You can use a GitHub account for that, or just download the Numbat folder without login, as it is free open source software.

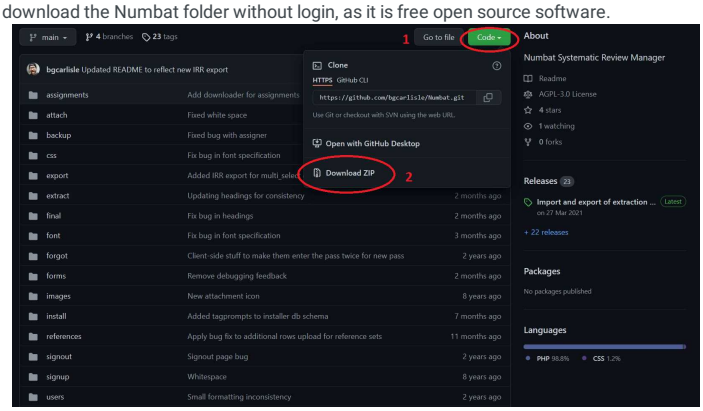

Fig. 4.1. Numbat repository clone from github

4.2 Install XAMPP program package, which includes Apache distribution and MySQL database.

XAMPP 8.1.5
Windows, Linux, OS X
by Apache Friends

This is what XAMPP looks like before any distributions are started:

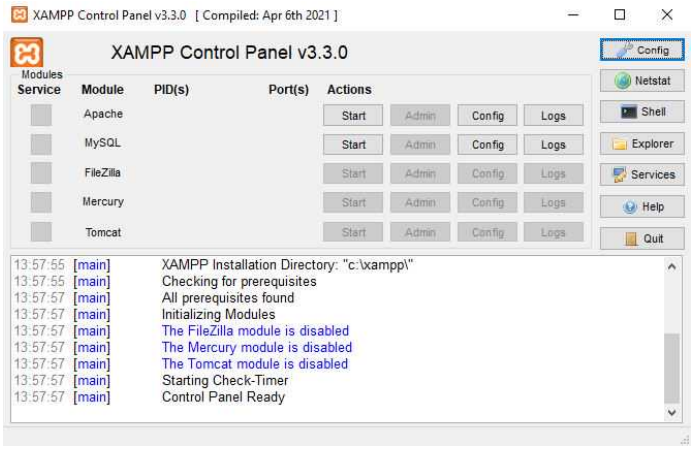

Fig. 4.2. XAMPP program (inactive mode)

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

4.3
- Copy the whole repository in the XAMPP folder **'htdocs'**.
- **In XAMPP**: Start the **Apache** module by clicking on '**Start**'
- Click on '**Admin'** button by Apache - you must see the installation instruction if it is still not completed.

The following image shows the successful installation:

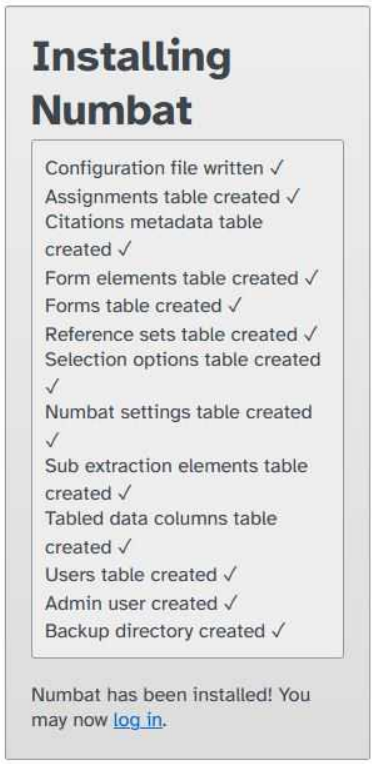

Fig. 4.3. Numbat installation via Apache admin

4.4 (probably provided by the company)
- **In XAMPP**: start the **MySQL** module by clicking on 'Start'
- Click on '**Admin**' by MySQL to go to the database
- Create a new **MySQL** *database* for Numbat

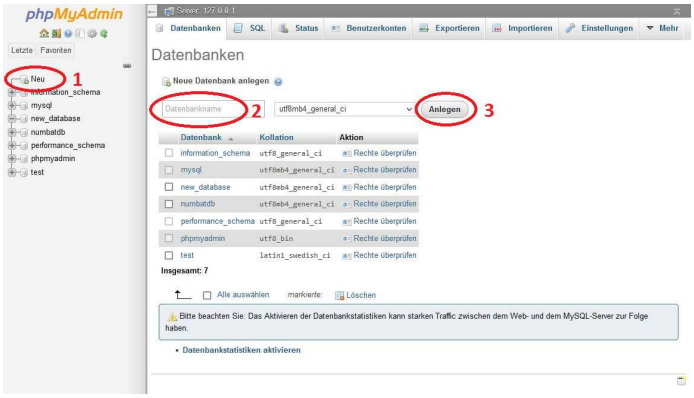

Fig. 4.4.1. New database in MySQL

- Create a new *user* (yourself as a new admin); you can use the default preferences with your name and password as shown in the figure 4.4.3

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

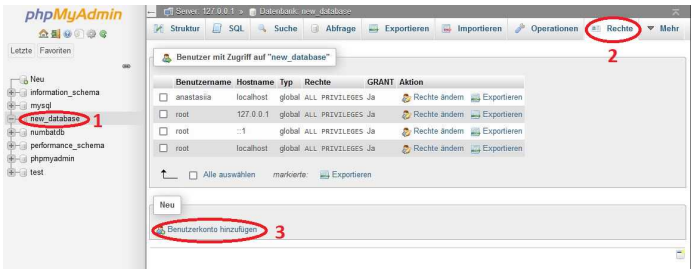

Fig. 4.4.2. New admin user in MySQL

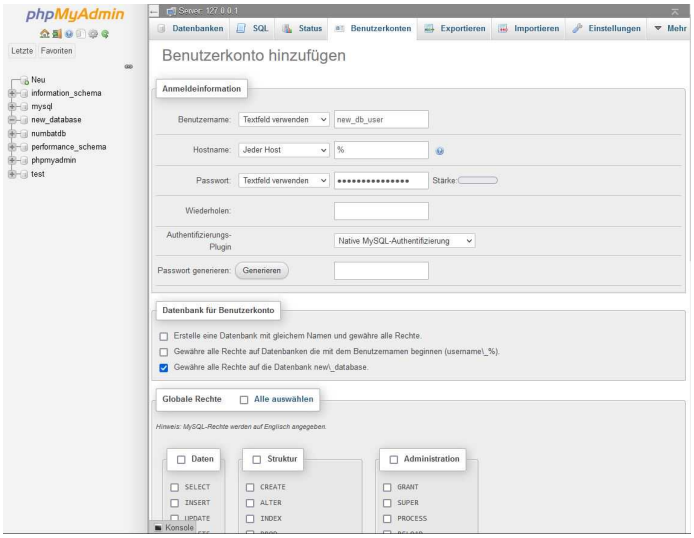

Fig. 4.4.3. Add user account in MySQL

4.5
- In XAMPP: use **'Admin'** button by Apache to complete the installation.
- In the opened window, fill out the *database username*, *password, name*, and *host*, as well as a *URL name* - everything but the URL you have just created in step 4.4.

Set up Numbat extraction workflow

5 **Preparation of the output from ODDPub for further evaluation in Numbat**

To validate the presence of Open Data underlying the article publication, an Extraction Form named Openness was created in Numbat. It follows the Open Data Criteria for LOM (*leistungsorientierte Mittelvergabe* or performance-based allocation of funds). Going one by one, it is checked for each publication, whether the Open Data statement detected by ODDPub ('is_open_data = TRUE') in fact refers to an openly accessible dataset.

A	B
User administration	Assignment of the account for new users
Manage reference sets	Uploading and editing lists of datasets (only text format such as .tsv allowed)
Edit extraction form	Implementation of extraction form
Attach files to references	Upload of documents to link to records (not relevant here)
Manage extraction assignments	Assignment of extraction forms AND / OR individual data records to specific users
Do extractions	Actual checking of publications for open data
Import extractions	Upload of further data to extract which was missing in the already uploaded dataset or was collected outside of Numbat
Reconcile finished extractions	Overview of completed datasets and merging of answers from several users
Export data	Export of finished table after test has been completed
Backup data	Create a backup of all information

Tab. 5. Menu items in Numbat and their descriptions

New user registration:

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

5.1 If you want to register as a **new user** to the existing Numbat instance, click on *'New here? Sign up'.* You need only your Email, password, and name to register. The Numbat admin will then activate your account.

User administration:

As the **admin** of the existing Numbat instance, go to the main menu item *'User administration.* New users will have unverified Email addresses. Verify it and assign *privileges* - User or Admin. The privileges are described on the same page.

5.2 To set up the Openness extraction form in your Numbat workspace, use the item ***Edit extraction form*** and go to ***Import an extraction form***. button The prepared extraction form should be in JSON format.

5.3 Filter the output from ODDPub in the ***is_open_data*** column for **TRUE** statements. You can use either table calculation programs like Excel or any text editor that you have on your computer.

5.4 Save these filtered data in **.tsv** format (tab-delimited text file) as a new input for Numbat. For that, a text editor is more appropriate than Excel.
Excel spreadsheets are known for causing unexpected errors, e.g. from adding a small symbol or a space, which can lead to incorrect file reading, although the spreadsheet might seem fine at visual inspection. The copy and paste actions are also dangerous in Excel, by which some data can be lost or modified. Using a text editor, you are in control of any characters.

6 Load the article DOIs and detected statements into Numbat

6.1 Select '**Manage reference sets**' in the Numbat menu, then '**Add new reference set**'.

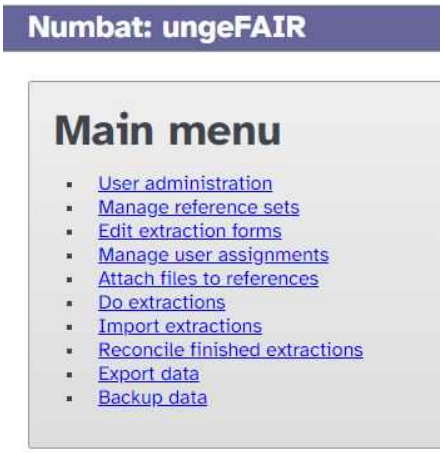

Fig. 6.1. Numbat main menu

6.2 Select the relevant columns from the dropdown menu (here: *doi, is_open_data, open_data_statement*) and give the set a name (e. g. 'Publications of 2020').

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

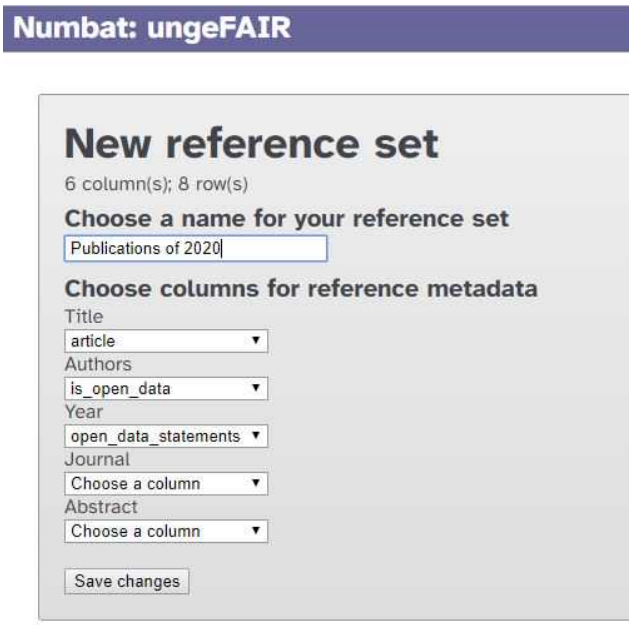

Fig. 6.2. Process of adding a new reference set

6.3 Assign the new dataset to a user (or multiple users) via '**Manage extraction assignments**' in the **Main menu**:

1. Choose all records or select some of them from a list.
2. Choose an extraction form in 'For the following form'
3. Choose user(s) in 'For the following user'.
4. Click 'Assign to user' - a list appears below, in which successful assignments appear green.

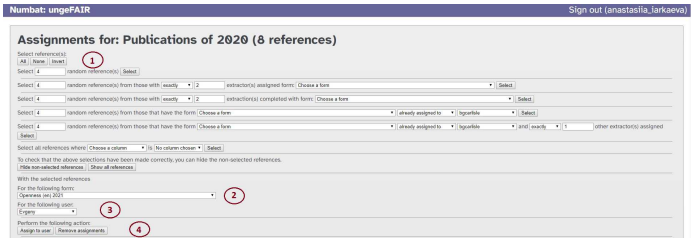

Fig. 6.3. Process of user assignment

Extraction with Numbat

7 Preparation for the manual extraction of information on Open Data status

7.1 Select the '**Do extractions**' option and start the extraction of Open Data status for publications where the ODDPub detected an Open Data statement (is_open_data = TRUE) by clicking on '**Extract**'.

7.2 Access the publication via DOI (http://doi.org/ + DOI from the list, e.g. http://doi.org/10.1128/mBio.02755-20).

A quick run of an example extraction, covering steps 7.2. through 8.18. is presented below. Please note, that while an actual extraction would have taken less time for this simple case, some cases require a lot more search and checking, and sometimes more than one dataset will be extracted per article (for this you would add a *new sub-extraction* and go through the evaluation of another dataset again).

In case there are multiple repositories, add each time a new sub-extraction after finishing the

previous one (see the video in Section 7.5).

7.3 From Open Data statements detected by ODDPub, get an overview of what is indicated regarding data sharing

Screen the detected statement(s) for e.g. information on the repository or accession codes. Also look out for false positives, which could be open code or statements completely unrelated to data sharing.

The extraction form includes not just 'Open Data' in the narrow sense (sharing of openly accessible data), but also restricted data access, as well as reuse of Open Data. It allows to document these two practices as well, without having to finish the full extraction, most items of which are then inapplicable. It is possible to save time by not extracting obvious cases of data reuse and restricted data access, but less obvious cases will appear during extraction, and then numbers on these practices would be incomplete.

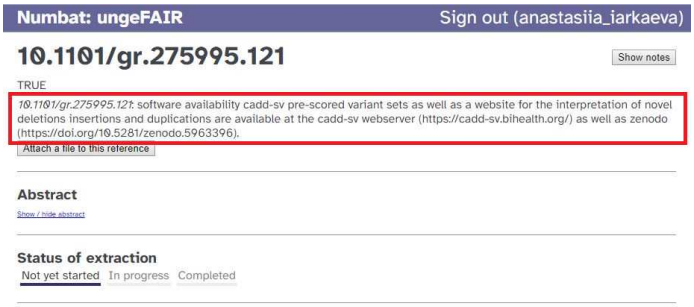

Fig. 7.3. Extraction of the dataset; Open Data Statement in the red frame

7.4 Search for other indications of Open Data in the article, missed by ODDPub

Despite high sensitivity, ODDPub can miss Open Data statements. This will be the more common, the more your field is removed from the biomedical field, for which this workflow was validated. It is also expected to become more common over time due to new repositories and standard statements. Thus, it is recommended (**but not strictly necessary**, depending on your use case), to briefly screen the article itself for further indications of shared datasets.

Take the following steps:

1. Search for keywords "*data availability*", "*data sharing*", and "*data access*"; if you thus find a section named "*data availability statement*" or similar, check the complete section for indications of shared data

2. If no such statement is found, proceed with the keywords "*dataset*", "*data set*", "*access**" and "*availab**", and check all hits, where there are <=10 for each keyword; where keywords yield >10 hits, dismiss them

3. If no such statement is found, search for the keyword "*data*"; if it yields <=10 hits, check each of them, if it yields >10 hits, dismiss

This procedure helps to detect a majority of datasets missed by ODDPub, with only a small time investment. However, for a larger set of articles, it is still substantial, so only implement this step if missing datasets would be detrimental to your use case.

As a compromise between sensitivity and effort, it is **recommended to always open the article**, **navigate to the statements** detected by ODDPub **and check their immediate surrounding** for information on datasets which might have been missed. For that, you can also use the keyword search in the article to quickly find the section(s) from the ODDPub statements. If the statement looks like a combination of different sections/sentences, use different keywords from the statement to access all detected sections.

7.5 Begin a new sub-extraction

For each repository extract one dataset.

If

1. Add a new **Sub-Extraction** to the current extraction.
2. If datasets have been shared in two or more repositories, start with the first repository listed.
3. For a given repository, if more than one dataset has been shared, choose the first listed for the extraction.
4. Repeat the extractions by adding a new sub-extraction for each new repository. In the end, there will be as many sub-extractions, as dataset repositories are named in the article (plus potentially supplemental data).
5. Typically, you will have one or two repositories and a few datasets per repository, but a larger number of each is possible.

Fig. 7.5. Starting the extraction by adding a new sub-extraction

How to add a new sub-extraction (and delete spurious sub-extractions):

Openness extraction form

8 Answer all the questions in the extraction form.

Use the open data definition shared here.
Beware that this definition includes data shared under restrictions. However, only information on openly shared data is fully extracted using this workflow.

Also note that the definition includes a few very specific requirements, necessary to align with our use case of incentivizing open data sharing at a research performing organization. Make sure to adapt the criteria not only to your institution (if applicable) but also to your specific use case. The adjustments will likely be small, but regard e.g. the time period for which open data are considered eligible, and by when possible embargo periods have to have expired.

The following is a short summary of these criteria, which is also accessible in the Openness extraction form under 'The requirements for Open Data (click (?) for more information)':

(i) The datasets are explicitly referred to in the publication; a reference e.g. to Supplementary Materials without further explanation is not sufficient.

(ii) The data are shared in a machine-readable format; for tables e. g. Excel format, for text e. g. TXT format; PDFs are conditionally machine-readable - they are suitable for texts (with sufficient tagging or structuring), but not for tables because they are not being recognized as such.

(iii) The data allow the analytical replication of at least one part of the study results and/or new analyses; listing statistical values (average, standard deviation, p-value, etc.) is not sufficient.

(iv) The data record can be retrieved online independently of the article. Tables embedded in articles do not count as open data unless they can be accessed as independent digital objects. For this, they have to have their own DOI or accession code AND have been shared independently in an external data repository or website.

Further explanations are to be found below and in the corresponding individual steps of the Numbat checking workflow.

A quick video tutorial on how to do the extractions can be found under Section 7.2.

8.1 Is there a clear reference to available datasets in the publication?

(i) Check the article type first, because article types such as 'review', 'opinion', or 'additional' are excluded from the examination, as they do not represent original research and no new datasets were generated as their basis. Meta-analyses and systematic reviews are considered as (potentially) generating new data and are extracted further. If the article is of a type which does not generate new data, answer 'inapplicable'.

(ii) There must be a statement either explicitly indicating that raw data have been shared or at least leaving it open, how raw data are. 'Raw' is defined loosely here but excludes statistical summary data (also see step 8). A general statement that the supplement contains 'additional information' without any further explanation is not considered a 'clear reference', and no further checking is necessary.

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

To assess this, check whether there is a data availability or data sharing statement/section in the article (see steps 7.3. and 7.4). References to supplemental data must be pursued further, whether the supplements were shared in a repository or on the journal/publisher website. In the latter case, a dataset identifier will be typically unavailable and thus does not need to be entered.

(iii) Data which is to be requested exclusively from the author or data contributor ("data available upon (reasonable) request") is not considered to be available within this extraction.

Fig. 8.1. Q1: Reference to dataset(s)

8.2 Can the data be found?

Check, whether the dataset can be found in the indicated location, and thus, whether the findability of the dataset itself independently of the article is given (supplements on journal/publisher websites do not comply with this, but this is checked in step 8.4). For this, use the identifiers, links, or repository names plus accession codes to open the landing page of the dataset. If the indicated dataset is not directly accessible in this way, search the indicated repository and/or use search engines, according to the below criteria.

Apply the following criteria to assess the findability of datasets:

(i) Mentioning the repository/database without indicating the accession number or how else the data are to be found in the database is not sufficient to consider the dataset findable.

(ii) "Private links" to datasets, typically intended to share data with reviewers, do not comply with the findability requirement, as the dataset is only accessible through the article.

(iii) Findability in repositories is given, if a search in the repository using the first and last author names and the first five words in the paper title yields a corresponding dataset amongst the first ten search results. If a specific dataset title is indicated, this should be used instead.

(iv) Findability via accession numbers and database names using search engines such as Google to find the database link is still compatible with Open Data. In many cases, the accession number contains the name of the repository (e.g. GSE145594, egas00001004772). The procedure for searching is as follows:

i. The accession number (and repository name, if available) is/are searched for in Google.

ii. If it is among the first 5 search results, the first link should be screened to see whether this is the requested dataset. Use overlap with the article regarding e.g. author names, title, or year.

(v) A link to the own website of the author/researcher or institution is not sufficient unless it is immediately and without further searching evident where on the website the dataset is located.

Fig. 8.2. Q2: Findability

8.3 Please state the identifier (preferably a link or DOI) of the data that will be used in this extraction.

Both for DOIs or other accession codes document an identifier in form of a link. A unique identifier (PID) such as a DOI, a link (URL), or an accession number (always with the database name in which the data is stored) is required for findability.

Please keep in mind that a DOI takes precedence over a URL, as it is the more persistent identifier. So if there is a DOI associated with the data, please reference it.

(i) If the dataset has a DOI (digital object identifier), enter the DOI in form of a link (e.g. https://doi.org/10.1109/XXXX).

(ii) If the dataset is referenced by database/repository name and accession code, enter the full link, which will typically contain both the database name and the accession code (e.g. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68849).

(iii) Furthermore, sometimes links in the article lead to the group of datasets where the actual data cannot be found or it is unclear which dataset was used in the study. The documented link should be for a page where the data can be actually downloaded or where the opportunity for this exists.

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

- The following video shows an example of a dataset that was linked in the article to the metadata record of the dataset (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD010370), but not the downloadable data. In this case, make sure to document the link with the data (https://www.ebi.ac.uk/pride/archive/projects/PXD010370 in the example).

(Video? - out of storage)

(iv) Only if a link cannot be given, enter the database and accession code separately.

(v) If supplements were shared in a repository or on the journal/publisher website, a dataset identifier will be typically unavailable, and thus does not need to be entered.

Fig. 8.3. Q3: Identifier(s)

8.4 Has the data been shared in a repository?

(i) Data in the supplement shared on the journal or publisher website meet the FAIR criteria to a particularly low degree and are difficult to find and reuse. If data have been deposited in this way, they are not considered compatible with Open Data requirements and no further assessment takes place.

(ii) Supplementary data deposited in repositories and findable through this repository, and thus independently from the article, are considered compatible with Open Data. A repository is here defined as any platform which allows to access data online, assigns identifiers or minimally a weblink to a dataset landing page, and offers metadata annotation to increase data findability. Acceptable repository types are general-purpose (e.g. Zenodo, figshare), field-specific (e.g. GEO), and institutional repositories (e.g. Harvard Dataverse).

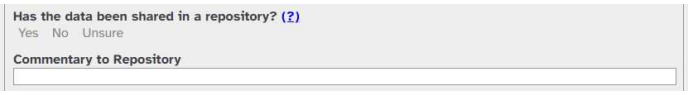

Fig. 8.4. Q3.1: Repository/supplements

8.5 Select the applicable repository name from the tag list

If the applicable repository is not on the list, then add a new tag.

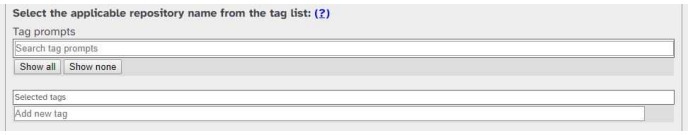

Fig. 8.5 Q3.2: Repository name

Following repositories are included in the current Openness extraction form version:

Addgene; ArrayExpress; BioProject; Clinical Proteomic Tumor Analysis Consortium Data Portal; Code Ocean; dcc; decipher; DRYAD; European Nucleotide Archive; figshare; FlowRepository; Gene Expression Omnibus; GigaDB; GIN; GitHub; Global Health Data Exchange; gpcrmd; Harvard Dataverse; Mass Spectrometry Interactive Virtual Environment; Mendeley Data; MetaboLights; NCBI Nucleotide; Open Science Framework; OpenNeuro; PeptideAtlas; PhysioNet; proteindiffraction; ProteomeXchange; PRoteomics IDEntifications Database; Sequence Read Archive; Synapse; The Electron Microscopy Data Bank;The European Genome-phenome Archive; TUdatalib; Worldwide Protein Data Bank; Zenodo

8.6 Enter the year of publication of the most recent dataset version (use only a year in the format YYYY):

If only one date is available, use this date.

Otherwise, use the year in which the latest major change to the dataset (upload or update) took place. This most often includes added data or new versions of old data. Changes to the record itself

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

(e.g. new keywords, saving in new formats for same data) are typically not to be considered major changes. However, if uncertain, be liberal and use the latest date of any change to the entry available in the dataset metadata.

Fig. 8.6. Year of the dataset publication

8.7 Can the data be accessed?

(i) Access to the data is possible, i.e. the data can actually be downloaded. This can typically be confirmed by opening the download window. If this is possible, it is to be assumed that data could be downloaded, and an actual download is not necessary. However, in case of doubt, download data to confirm that the downloadable files are actually data and not merely the corresponding metadata.

(ii) If the data are available under access restrictions (including any type of registration or confirmation of terms of use), they are not to be considered as Open Data. (For our institutional incentives program we do, however, consider restricted-access data, and assess them independently outside this extraction workflow.)

(iii) If the data are unavailable due to server downtime or a server error assumed to be temporary, data accessibility is to be checked again on a later day. Unavailability on two different dates is not compatible with Open Data status.

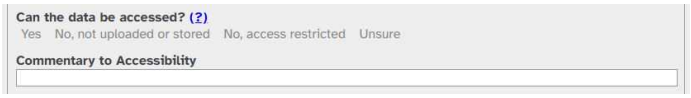

Fig. 8.7. Q4: Accessibility

8.8 Was the dataset shared under a standardized license?

A license gives the user information on how the dataset can be re-used (or not re-used at all). Often the data is not licensed in a standardized way, and instead, there is a 'bespoke license', which is stored only in one place in the repository and applies to each individual record.

For this question look for a standardized license such as Creative Commons (CC BY, CC BY-SA, etc.) or Open Data Commons (PDDL, ODbL, etc.) on the dataset landing page.

Fig. 8.8. Q4.1: License

8.9 Select the applicable license name from the tag list:

If the license displayed on the record landing page is not in the list of tags, add a new license tag in the corresponding window.

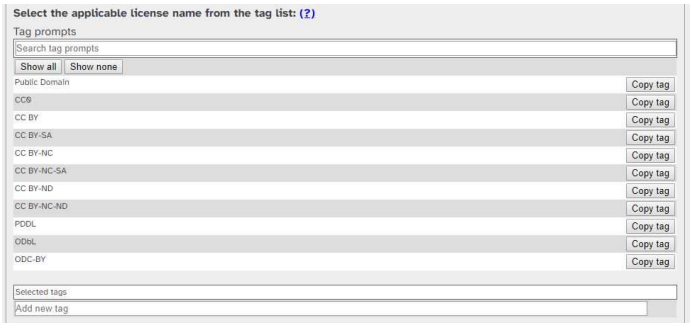

Fig. 8.9. Q4.2: Known licenses

8.10 Has the shared data been generated by the authors of the corresponding article („Own Data“) or is it re-used data generated by others („Data Reuse“)?

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

(i) We classify as "data reuse" all cases where the detected statement indicates that data were collected by others than the authors and were made available through a repository (openly or via a restricted but standardized access route). However, none of the further openness criteria (steps 7.10 to 7.13) is checked for cases of data reuse.

(ii) If at least one author of the article is also a contributor to the dataset, it is considered to be Own Data.

(iii) In case no contributors are indicated in the dataset metadata, an explicit reference about Own Data sharing in the publication is sufficient to assume that the data are newly generated data.

(iv) Within the extraction, there is no restriction applied as to whether the own dataset has already been the basis of other publications, or how much time passed between the publication of the dataset and article. (For our allocation of incentives we do, however, manually exclude datasets which were shared over three years before article publication).

(v) Indicating both Own Data and Data Reuse for the same article is NOT possible. In this case, give preference to the classification as Own Data.

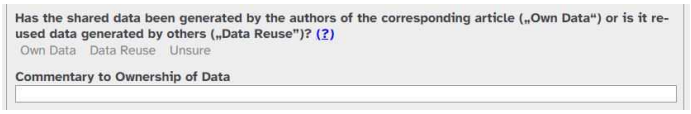

Fig. 8.10. Q5: Own/reuse data

8.11 Has the data been shared in a machine-readable format?

For tables, the machine-readable formats are e.g. CSV or Excel files, for texts –TXT, DOC, unformatted text, or XML.

PDF and PDF/A are conditionally machine-readable, but because there is no adequate structuring in this format type and it is resource-intensive to check the exact format of a PDF file, PDFs are not considered machine-readable here. Other types of files are always entered as machine-readable. The format might be undetermined e.g. when data are packed in zip files. If these are >200 MB large and a download would be required to check the file type, the format is indicated as undetermined. File formats are typically undetermined for rare data types or very large data, and then it is assumed that the files are in line with Open Data criteria, and the extraction is continued.

Lists from ETH Zurich and Publisso can provide support in case of doubt regarding the machine-readability of file formats, but it should be noted that machine-readability is not the same as suitability for digital long-term archiving.

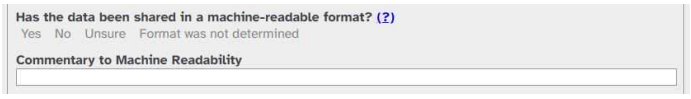

Fig. 8.11. Q6: Machine-readability

8.12 Which format is the data presented in?

Indicate data formats by clicking on the respective tags. Multiple entries are possible.

Fig. 8.12. Q7: Formats

8.13 If the data is image or audiovisual data: does the data have more than just illustrative character?

Data are purely illustrative if they serve as examples and do not form the full basis for at least a part of the analysis presented in the publication (e.g. for one figure). However, in individual cases, it can be difficult to evaluate whether this is the case. If image or audiovisual data stored in repositories are not explicitly described as examples and if there are at least three files, in case of doubt it can be assumed that they are not purely illustrative.

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

Fig. 8.13. Q8: Illustrative character of data

8.14 Does the data allow the analytical replication of at least some results?

(i) Data must be primary, unprocessed, or "raw" to the extent that they can be the basis for analyses presented in the publication. Thus, tables with individual data point values from which article figures were derived, can be compatible with Open Data, as long as these values are original or preprocessed (e.g. corrected, normalized) outcomes. Indices or other highly processed values, as well as summary statistics, are, in contrast, not considered compatible with Open Data. In line with this, lists of genes, proteins, or similar units with associated individual p-values are not sufficient. The pure amount of data is not decisive. Even very small tables can contain raw data if the sample size is small. If several measured values are listed for a unit of observation, sufficiently unprocessed data can be assumed in case of doubt. Generally speaking, borderline cases inevitably occur at this step, in which case we tend to be liberal.

(ii) OMICS data in disciplinary formats – given they fulfilled all criteria so far - are always to be assessed as allowing analytical replication. (This criterion was introduced partly for lack of disciplinary expertise, and could be adjusted, if such expertise is present.)

To allow for granularity in the decision process, "yes" and "no" decisions are both split into cases of either high or low certainty. The latter is indicated by the reservation that the choice "tends to be" yes or no. In either case the extraction continues with step 8.15., in which an overall assessment of Open Data status is performed.

Fig. 8.14. Q9: Analytical replication

8.15 Have the Open Data requirements been met? Is a discussion necessary?

Enter the overall assessment of the Open Data status of the extracted dataset. We recommend extensive commenting, especially at this step, to make sure that difficult decisions can be reconstructed. Please note that, comments made here are, however, not available to other extractors if the same dataset is additionally assigned to them.

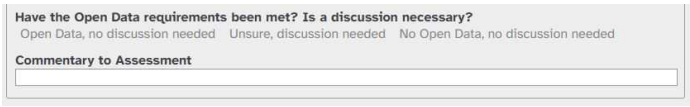

Fig. 8.15. Q10: Overall Open Data assessment

8.16 Reference to stop the extraction:

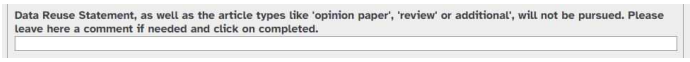

Fig. 8.16. Q11: Stop the extraction for some cases

8.17 Add a new dataset by adding a new sub-extraction:

If the article refers to data from further location(s) (typically if more than one repository is mentioned per article), click on "Add new sub-extraction" to begin the next extraction of Open Data status for a dataset in a different location.

Fig. 8.16. Q12: Reference to the further dataset(s)

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

8.18 Be sure to click on **"Completed"** (back at the top of the page) after finishing the extraction!

The output in form of a .csv file can be downloaded any time from the Numbat server, where all data about criteria decisions, as well as the final decision are stored.

Post-process extracted table

9 How to **export the final report** table:

The export is only possible if more than 2 data records have been checked.

Before the results table is downloaded, the answers must be reconciled (see above) if data records have been checked by several users, otherwise, there may be duplicates.

1. Select the '**Export data**' area -> '**Export Openness extractions**' -> download the results table.
2. Clean up the table:
   - For Excel: choose 'Data', then 'Text in Columns' -> Select 'Delimited' + tab stop / comma (depending on) + Standard -> Finish. Save in needed format.

Consider that the extraction dataset can contain several Open data assessments per article. If an analysis on the article level is desired, process the output in the desired way. In our case, for incentivizing data sharing on the article level, we detect all publications in which at least one dataset was shared (openly or with restricted access).

The output in form of a .csv file can be downloaded any time from the Numbat server, where all data about criteria decisions, as well as the final decision are stored.

article	is_open_data	open_data_category	is_open_code	open_data_statements	open_code_statements	reference_to_da
10.1038/s41467-020-16734-3	TRUE	general-purpose repository	TRUE	code availability all code used to analyze the dataset is openly available within lead-dbs/-connectome software (https://github.com/leaddbs/leaddbs).	code availability all code used to analyze the dataset is openly available within lead-dbs/-connectome software (https://github.com/leaddbs/leaddbs).	yes
10.1038/s41467-020-16929-8	TRUE	field-specific repository	FALSE	proteomics data have been deposited to pride server under accession code pxd017341 [http://proteomecentral.proteomexchange.org/cgi/getdataset?	NA	yes
10.1186/s12916-020-01851-z	TRUE	supplement	FALSE	additional file 5. kaplan-meier raw data	NA	yes

Tab. 9. Expected output table

Different issue handling in Numbat

10 The extraction up to this point was linear. Steps 8.1 to 8.4 describe steps which are optional. Sometimes, an errors needs to be corrected (8.1), or one or several dataset(s) need to be added to the extraction list. Often, the case occurs that an extractor is "unsure" about the Open Data status of a dataset. In this case, extractions can be assigned to another extractor (8.3). Assessments of two (or more) extractors from either unsure cases or instances where extraction by several extractors is desired for quality assurance purposes, can then be reconciled (8.4).

10.1 In case an extraction has to be **corrected**:

1. Go to the section '**Do extractions**'.
2. Find the dataset you want to correct.
3. Open the extraction form as usual (click '*Extract*').
4. Change the answer(s) as in a normal workflow and if needed click on '*Completed*'.
5. If the extractions have been completed and export has already taken place, delete the older version of the table from your storage and export the updated output via **'Export data'.**

Note: all answers following the affected answer downstream i.e. are logically related to that answer, will be **deleted**. It is made to avoid the internal inconsistency. In that case, the extraction has to be re-done for all questions following the affected one.

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1

E.g., as time passed and you change your mind about the first question, whether there is a clear reference to the dataset in the article, from 'yes' to 'no', you will skip all the following questions except the last one. The previous answers will be overwritten after you click 'completed'.

10.2 How to **upload new article** to the existing reference set (table):

1. Create a new document in the .tsv format locally in your text editor, which contains the same column names as your primary existing set, as well as all references (DOIs) that must be added.
2. Add new reference(s) under '**Manage reference set'** --> '**Your set name XY'** --> **"'new reference'**.
3. Check the matching of columns in the updated table.
4. Assign a new record to one or several user(s) via '*Manage user assignments*'.

10.3 How to **assign a dataset** in question to another extractor:

There is an '**Assign to'** option at the end of every extraction. Select the extractor(s) and the extraction form (here: *Openness*), then complete the test by clicking on '*completed*'.

For our use case, we assigned datasets to other extractors always if Open Data status was 'unsure', and in turn never did so, if it was clear to extractor #1 (i.e., 'yes' or 'no').

10.4 How to **reconcile** the answers of different extractors:

This function is needed when the same articles are extracted by several persons. If more than one extractor has done the extraction, the answers are compared and then reconciled i.e. combined. The overlapping responses can be reconsiled right away; the different responses can be discussed further or the decision is made according to the commentaries of the extractors.

1. Select **'Reconcile finished extractions'** area.
2. All extractors who have completed the extractions are shown in the column '*Extractors*'.
3. Click on '**Reconcile extractions**':
   - all **green** areas are consistent. You can skip these responses.
   - the **red** areas must be compared - for this, a discussion amongst the extractors may be necessary to come to an unambiguous conclusion, unless one person is considered a 'master extractor', and undertakes a final assessment by him or herself. Pay attention to the commentaries as they can include useful information to make a decision.
4. There are lines with answers of different extractors and the last line called '**Final copy'** in every question.
5. Save the selected answer in the final report by clicking either **'Copy to final'** or selecting the answer in the final copy line.
6. Continue until all red areas become green.
7. As in the article extraction, don't forget to click '*completed*' after finishing reconciliation.

Citation: Anastasiia Iarkaeva, Evgeny Bobrov, Jan Taubitz, Benjamin Gregory Carlisle, Nico Riedel Semi-automated extraction of information on open datasets mentioned in articles https://dx.doi.org/10.17504/protocols.io.q26g74p39gwz/v1