Jul 15, 2025 Version 2

🌐 Semi-automated extraction of information on open datasets mentioned in articles V.2

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

Evgeny Bobrov
BIH at Charité

Protocol status: Working
We use this protocol and it's working

Abstract

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. Operationalization of Open Data criteria has been described in the article preprint **Operationalizing Open Data – Formulating verifiable criteria for the openness of datasets mentioned in biomedical research articles** (Bobrov et al., 2023).

Materials

- 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**
- Hardware specifications:
1. CPU: standard modern CPU, e.g. Intel i5 or equivalent
2. Memory (RAM): minimum 8 GB, recommended 16 GB
3. Storage: 100 GB free storage

Troubleshooting

Automated validation of publications for statements indicating Open Data using ODDPub

1 **Collate a list of article identifiers**

Begin with a list of articles identifiers (usually DOIs), for which you want to assess the openness of the datasets underlying these articles. These idetifiers can typically be obtained by **searching publication databases** (e.g. Pubmed, Web of Science, OpenAlex, Dimensions, Embase) using the relevant search criteria and exporting the results with the relevant metadata fields.

Search criteria may include *institutional affiliations of authors*, specific *research fields*, specific *journals*, and typically, a *publication date range*. If the focus is on articles from a particular institution, a **curated list of publications** might be available directly from the institution (often provided by the institutional library).

The development and optimization of ODDPub (Open Data Detection in Publications) algorithm have been aligned with Open Data criteria, similar to those cited below. However, some changes in the criteria have been introduced over time.

Citation
Evgeny Bobrov, Nico Riedel, Miriam Kip (Invalid date) . Operationalizing open and restricted-access data—Formulating verifiable criteria for the openness of data sets mentioned in biomedical research articles. Quantitative Science Studies, 2024; 5 (2): 383–407. 10.1162/qss_a_00301   LINK

2 **Obtain the article full texts**

There are several options for this:
- via **PubMed Central** (Open Access articles)
- via **unpaywall API**: the full-text links (Open Access articles).
- via Publisher APIs: for full-text retrieval of subscription-based articles, offered by several major publishers, like **Elsevier**, **Wiley**, or **Springer/Nature**
- via **full-text R package**: a solution that combines several of those data sources for downloading full texts.

These retrieval options are limited to articles that can be accessed as Open Access or through the subscription provided by the institution where the article are retrieved.

Store all retrieved article full-texts in a single folder either as PDFs (preferred for current ODDPub version) or as text files (e.g. XML).

3 **Apply ODDPub to article full texts**

> **Note**
>
> <u>ODDPub</u> (Open Data Detection in Publications) is an open source text mining algorithm implemented in R. It screens articles for data sharing statements throughout the article, using keywords and keyword combinations. To use ODDPub, the publications must be prepared in PDF or text file format and stored in a local folder. Only full-text publications can screened. There is no limit on the number of publications.
>
> The ODDPub workflow involves **three steps**:
>   1. Generation of **text files** out of PDF files (only if the full text is not already available in a text format, e.g. XML).
>   2. Search for the **data availability statements** (DAS) and **keywords** (in and outside of DAS), defined in the script, such as *repository names, accession-identifier-similar strings, pre-defined data sharing expressions,* or *references to supplementary material.* Keywords can be found both within and across sentence boundaries, potentially spread throughout a paragraph.
>   3. Matching detected **keywords and regular expressions** from the script. If a keyword group is matched, the publication is categorized as containing Open Data.
>
> In addition to detecting Open Data statements, the algorithm detects the location of shared data (*general-purpose repository, field-specific repository, supplement, data jorunal*), as well as statements related to Open Code (*open source software*).
>
> Recent updates to ODDPub (v7.0.0) introduce distinctions between **(own) Open Data** (column "is_open_data") and **Data Reuse** (column "is_reuse") categories, along with new categories like *"upon request", "github", and "unknown/misspecified url"*. Detected *Data Availability* and *Code Availability statements* are stored in separate columns.
>
> For more information about ODDPub's implementation, see the following sources:
>
> 1. Development, functionality, validation, and performance of ODDPub:
>
> > **Citation**
> >
> > 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   LINK
>
> 2. Performance, scalability, and outlook of ODDPub:
>
> > **Citation**
> >
> > Iarkaeva, A., Nachev, V., & Bobrov, E. (2023)
> > . Workflow for detecting biomedical articles with underlying open and restricted-access datasets.
> > MetaArXiv.
> >
> > 10.31222/osf.io/z4bkf   LINK

The following steps describe how to use the ODDPub package in R.

3.1   First, install ODDPub package in R with the following command:

> **Command**
>
> Use e.g. RStudio. (ODDPub Version 5)
>
> Installation of ODDPub
>
> ```
> # install.packages("devtools") # if devtools currently not installed
> devtools::install_github("quest-bih/oddpub")
> ```

3.2 Next, set the working directory to the folder containing all the publications to be examined. In this example, the folder is named 'PDFs':

Command

R command for setting working directory

R command for setting working directory

```
setwd("C:\Users\username\path\to\PDFs")
```

Create a new folder to store the converted text files from PDFs after algorithm application.

3.3 After installing ODDPub and organizing the PDFs, run the following command to start evaluation for Openness:

Command

A minimal R-Script to run the Open Data and Open Code detection via the ODDPub algorithm (ODDPub Version 7.0.0)

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)

#  Keep only actual DOIs in the data frame
oddpub_results$doi <- oddpub_results$doi %>%
str_remove(fixed(".txt")) %>%
str_replace_all(fixed("+"), "/")

# Write results in CSV file
write_csv(oddpub_results, "oddpub_results.csv")
```

After screening, an **output** file in *.csv* format will be created in the same folder as the ODDPub script. You can change the format to any other type if needed (e.g. .txt, .tsv). Below is an example of the output from the latest ODDPub version (7.0.0). This output will be further validated manually in Numbat (see from the step 4).

Expected result

	doi	is_open_data	open_data_category	is_reuse	is_open_code	das	open_data_statements	cas	open_code_statements
	10.1002/alz.12763	FALSE	re-use	TRUE	FALSE		all fdg-pet scans used in this study were downloaded from the adni server in fully pre-processed format (see http://adni.loni.usc.edu/methods/documents/ for details) and then spatially normalized to a customized fdg-pet template in montreal neurological institute (mni) standard space using spm8.; data used in the preparation of this article were obtained from the adni database (http://adni.loni.usc.edu/).		
	10.1002/14651858.cd014963.pub2	TRUE	general-purpose repository, upon request	FALSE	FALSE		the completed rob 2 tool with responses to all assessed signalling questions is available online at: https://zenodo.org/record/6500842.; in an attempt to address these issues and since there is not yet an established tool for critically appraising platform trials we pioneered a checklist (park 2020) with results available at https://zenodo.org/record/7015269#		
	10.1002/14651858.cd011740.pub2	TRUE	general-purpose repository	FALSE	FALSE		the data are available in the open science framework (osf.io/7tydm/).		
	10.1001/jamanetworkopen.2022.1875	FALSE	upon request	FALSE	FALSE	data sharing statement: anonymized data will be made available to the scientific com			

doi	is_open_data	open_data_category	is_reuse	is_open_code	das	open_data_statements	cas	open_code_statements
					munity upon reasonable			
10.1016/j.ebiom.2022.104293 10.1002/jlb.2a0421-200r	TRUE	field-specific repository, general-purpose repository	FALSE	TRUE	data sharing statement with publication all deidentified ms proteomics will be openly available via the proteomexchange consortium (http://proteomecentral.proteomexchange.org) and the pride partner repository 18 with identifier pxd036590. all cov	data sharing statement with publication all deidentified ms proteomics will be openly available via the proteomexchange consortium (http://proteomecentral.proteomexchange.org) and the pride partner repository 18 with identifier pxd036590. source code and test data are available via github (https://github.com/gcaptur/ covid-proteomics).		source code and test data are available via github (https://github.com/gcaptur/ covid-proteomics).

doi	is_open_data	open_data_category	is_reuse	is_open_code	das	open_data_statements	cas	open_code_statements
					ids ortium hcw clinico demographic data including study protocols and templates of informed consent forms used for the study are freely available following a data access request through the covids ortium data access portal. source code and test data are available via github			

doi	is_open_data	open_data_category	is_reuse	is_open_code	das	open_data_statements	cas	open_code_statements
					(https://github.com/gcaptur/covid-proteomics).			
10.1016/j.waojou.2022.100703	FALSE	supplement, upon request	FALSE	FALSE	availability of data and materials the data used and analyzed for this study is available from the corresponding author on reasonable request.	table 1. demographic data at the date of start with mepolizumab therapy. all data		

Tab. 3.3. Example result after running ODDPub Version 7.0.0

Overview of extracting form in Numbat

4 The attached file contains the **Numbat extraction form** (Version 4). This form was automatically generated from Numbat in markdown format and then transformed into a PDF using **https://md2pdf.netlify.app/** for better human readability. It contains all questions related to Open Data criteria for manual openness verification.

_____
_____

📄 Openness form 2024.pdf 474KB

_____
_____

4.1 The table below provides an overview of all Open Data criteria presented in the extraction form.

question	answers
Is there a clear reference to available datasets in the publication?	Yes │ No │ Inapplicable │ Unsure
Is the detected reference found in the data availability statement (DAS)?	Yes, in a DAS │ No, and there is no DAS in the article │ No, but there is a DAS (= reference present, but outside of DAS) │ Unsure
Can the data be found?	Yes │ No │ Unsure │ Not checked (reuse)
Has the data been shared in a repository?	Yes │ No │ Unsure │ Not checked (reuse)
Please state the identifier (preferably a link or DOI) of the data that will be used in this extraction.	*open text field*
Select the applicable repository name from the tag list	*tags list + open text filed for new tags*
Can the data be accessed?	Yes │ No, not persistent │ No, access restricted - academic data │ No, access restricted - pharma data │ Data under embargo │ Not checked (reuse) │ No, not uploaded │ Unsure
Has the restricted data been generated by the authors of the corresponding article („Own Data") or is it re-used data generated by others („Data Reuse")?	Own restricted data │ Reuse restricted data
Enter the year of publication of the most recent dataset version (use only a year in the format YYYY):	*open text field*
Was the dataset shared under a standardized license?	Yes │ No │ Unsure │ Not checked (reuse)
Select the applicable license name from the tag list	*tags list + open text filed for new tags*
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")?	Own data │ Data reuse │ Unsure
Has the data been shared in a machine-readable format?	Yes │ No │ Unsure │ Format not defined
Which format is the data presented in?	XLS/XLSX │ CSV/TSV │ TXT/DOCS │ Other text or table formats │ Video │ Audio │ Image │ FASTA/FASTQ │ RAW │ Other generic format │ Other subject specific format │ Unsure
If the data is image or audiovisual data: does the data have more than just illustrative character?	Yes │ No │ Inapplicable │ Unsure
Does the data allow the analytical replication of at least some results?	Yes │ No │ Tends to be positive │ Tends to be negative
Have the Open Data requirements been met? Is a discussion necessary?	Open Data, no discussion needed │ Unsure, discussion needed │ No open data, no discussion needed

Tab. 4.1. Inquiries and Responses Related to Open Data Criteria in the Openness Extraction Form (2024)

A quick **video tutorial on how to do the extractions** can be found under Section 5.1.

4.2 Which types of articles might produce data and therefore should not be excluded from the screening as potential sources of data?
- case report
- study protocol
- methodology
- short reports
- systematic review

4.3 **Systematic reviews:**
- should include more than just a list of referenced publications; they should also provide extracted information and clearly outline the exclusion and inclusion criteria.

4.4 **What does not count as a 'clear reference'?**
- according to the current criteria (Bobrov, et al. 2024, **10.1162/qss_a_00301**) supplements are not considered Open Data, partly because they are not shared independantly from the article.
- reference to data shared within the manuscript, without explicitly mentioning the word 'supplement', are also not sufficient.
- references to data presented in tables within articles do not qualify as clear references to raw data.
- reference to data upon reuqest by the authors is do not qualify as clear references to raw data.

4.5 **What does count to the 'best identifier'?**
- the best option for an identifier is a *persistent* one, such as a DOI or Handle.
- if both a URL and a DOI are available, the DOI should be documented as it is a more stable and persistent identifier.
- for discipline-specific repositories that do not provide persistent identifiers, documenting the URL along with the accession number is the best approach.
- in cases like NCBI BioProject/SRA, link the project main page along with the corresponding SRA page leading to the data files.
- non-persistent URLs, such as those from GitHub or the Open Science Framework without a DOI, are not sufficient as identifiers or data sources because not just metadata, but also data files can be changed by the data owners.
- supplements are generally considered insufficient unless they are stored independently of the article.

4.6 **Handling the documentation of the data set publication date:**
- typically, use the last update date.
- however, this can be confusing in some cases, as the last metadata update date may not correspond to the publication date. For example, in the Gene Expression Omnibus (GEO), the "*Status (public)*", "*Submission date*" and "*Last update date*" are usually three different dates. The entry most closely related to the publication date is the "*Status (public)*".

4.7 **Managing Access with Registration Requirements:**
- registration in a repository or databank should not be mandatory to access the data. If access is hindered in such way, it cannot be considered open access.
- simply clicking on an agreement without requiring registration is sufficient to meet the definition of open and free access.

4.8 **Genetic sequences and case reports:**
- considered (open) data only when shared in a discipline-specific repository.

4.9 **Embargoed data:**
- data that are not yet public but whose publication is planned and precisely specified on the existing dataset landing page.
- they are documented through the "*Data under embargo*" button and can be validated after the embargo period ends.

4.10 **Pharma data:**
- researchers may collaborate with or be part of a pharmaceutical company, but the primary consideration is whether the data is stored on an academic platform/database/repository or at a pharmaceutical company (e.g., Vivli, Yoda).
- mention of pharmaceutical platforms in the article indicates that the data are more likely to be stored outside academic repositories, and such cases are treated separately from common (academic) open data cases.

4.11 **Criteria for restricted or reuse cases:**
- criteria for repository requirements and accessibility may vary for restricted data sets. A URL to the location or page where the data can be requested (rather than a general "*upon request*" statement in the article) is sufficient.
- the same approach applies to the reuse of restricted data sets.

4.12 **Own or Reuse data:**
- document both "Own (Open) Data" and "Reused (Open) Data" cases.

- typically, dataset metadata include author names, which can be compared with the article's authors to determine data ownership. In some cases, such as some of the NCBI repositories, the data owner may be listed only under an institution name. In such cases, if any author is affiliated with that institution, the dataset is considered "own" data.
- if neither author nor institution names are available, the article text (e.g., "*data were collected during...*") and context should help to clarify the data origin.
- reuse statements are often found outside the "Data Availability Statement".
- *reuse of the datasets generated by other researchers is not considered Open Data per default, but need a separate verification for openness.*

4.13   Are genomics summary statistics the same as other summary statistics?
- generally, summary statistics are not considered raw or unprocessed data. However, genomics summary statistics stored in specialized repositories, such as the GWAS Catalog, differ from typical summary statistics. They represent a collection of studies and are not directly comparable to standard statistics.

4.14   Are classifieds (models) for ML training raw data?
- classifiers for machine learning algorithms are not considered raw data according to our definition.

Manual validation using Numbat - User access

**5   Preparation for manual extraction of Open Data status**

For a detailed walkthrough of the manual extraction process of Open Data, see the instruction below under 5.1-5.4.

To access the publication, use the DOI directly on *doi.org* website or enter the URL **http://doi.org/** + DOI into browser's address bar, e.g. http://doi.org/10.1128/mBio.02755-20.

Please note that while this example may be straightforward, actual extraction can be time-consuming. Some cases may require extensive searching and checking, and occasionally more than one dataset may need to be extracted per article.

5.1   Select the 'Do extractions' option to begin extracting the Open Data status for publications where ODDPub has detected an Open Data statement ("is_open_data" = TRUE) by clicking on **'Extract'**.

If there are multiple repositories, add a new sub-extraction for each repository after finishing the previous one.
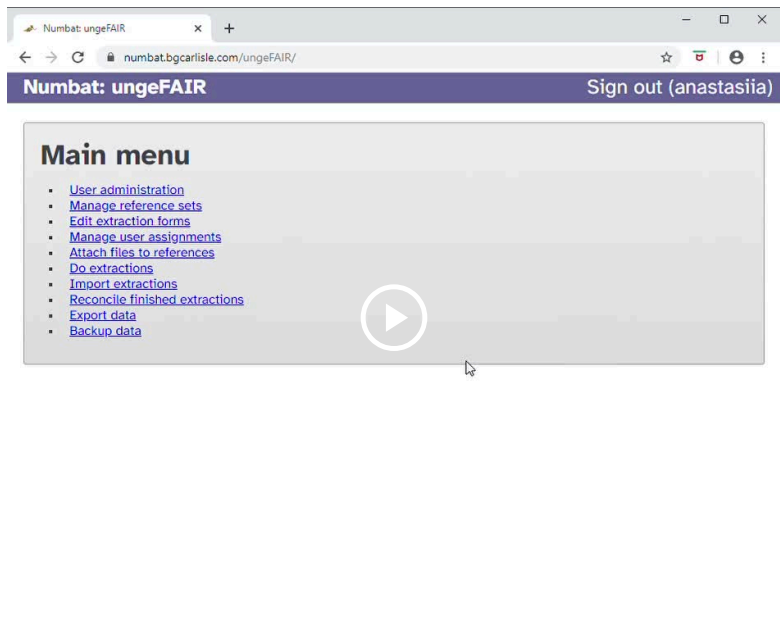
The extraction of other categories, such as "is_reuse" =TRUE or "unknown url", is optional and
should be based on the specific goals of the screening.

5.2   **Open Data statement - Which 'candidates' are detected by ODDPub?**

Review the detected statement(s) for details such as repository information or accession codes. Be sware of potential false positives, which may include open code or statements completely unrelated to data sharing.

The extraction form not only covers **open data** in the strict sense (freely accessible data) but also data under **restricted access** and the **reuse** of open data. It allows documentation of these practices without completing a full extraction.

Some **reuse cases** might lack a direct citation or referenceto the landing page in the article, even though they are indicated as a basis for the study provided by another institution or found in publicly accessible repositories. Where it is possible, answer all questions up to *authorship*. For cases where multiple criteria are challenging to assess, it is possible to document the **reuse cases** with the response *"Not checked (reuse)"*.

To save time, you can skip extracting cases where it is clear that data have been reused or access is restricted. However, in many cases, this clarity emerges only during the detection process. Ignoring obvious cases may result in an incomplete overview. picture will not be complete.

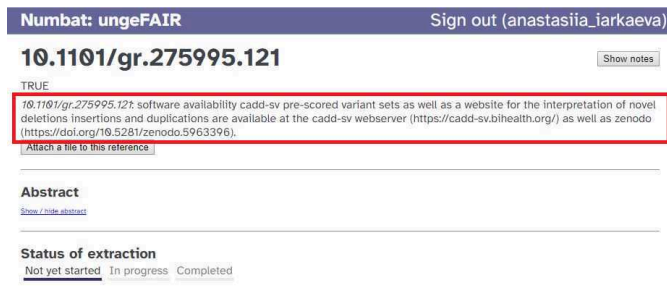Always click 'C*omplete'* to finish the extraction.

Fig. 5.2. Open Data Statement in the red frame

5.3 Search for other indications of Open Data missed by ODDPub

Although ODDPub is highly sensitive, it may still miss some Open Data statements, especially in fields less related to the biomedical field for which the workflow was developed and validated. To balance sensitivity and effort, it is **recommended to review the article**, focusing on the areas around **the statements** detected by ODDPub to **check for any additional** information that might have been missed. You can use keyword searches within the article to quickly locate the sections related to the ODDPub statements. If a statement appears to be a combination of different sections or sentences, use various keywords from the statement to access all detected sections.

As new repositories and standard statements emerge, the frequency of missed Open Data statements may increase over time. Therefore, it is recommended (**though not strictly necessary**, depending on your use case) to briefly scan the article itself for further indications of shared datasets. Follow these steps:

1. Search for keywords such as "*data availability*", "*data sharing*", and "*data access*". If you find a section named "*data availability statement*" or similar, review the entire section for indications on shared data.

 2. If no such section is found, use keywords like "*dataset*", "*data set*", "*access**" and "*availab**". Check all results with <=10 hits for each keyword. Dismiss results where there are >10 hits.

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

This procedure helps identify most datasets missed by ODDPub with minimal additional time investment. However, for a larger set of articles, this step may still be substantial, so consider implementing it only if missing datasets would significantly impact your use case.

5.4 Begin a new sub-extraction

For each repository extract <u>one</u> dataset.

1. Add a new **Sub-Extraction** to the current extraction.
2. If datasets are shared across multiple repositories, begin with **the first repository listed in the article**.
3. For each repository, if multiple dataset are shared, choose **the first listed dataset** for the extraction.
4. Repeat the extractions by adding a **new sub-extraction for each additional repository**. The total number of sub-extractions will correspond to the number of data repositories mentioned in the article, plus any additional sub-extractions for explicitly referenced supplemental data.
5. Typically, you will encounter one or two repositories and a few datasets each, but a larger number is possible.

Fig. 5.4. Starting the extraction by adding a new sub-extraction

How to add a new sub-extraction and delete any spurious ones:

Set up extraction workflow in Numbat - Admin access

6   **Preparing the ODDPub output for further evaluation in Numbat**

To validate the presence of Open Data associated with the article, an extraction form was created in Numbat. This follows the **Open Data Criteria** for LOM (*leistungsorientierte Mittelvergabe* or performance-based allocation of funds). Each publication is reviews individually to determine whether the Open Data statement detected by ODDPub ("is_open_data" = TRUE) indeed 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. 6. Menu items in Numbat and their descriptions

6.1   **New user registration:**

To register as a **new user** in the existing Numbat instance, click on *'New here? Sign up'*. You only need to provide your email address, password, and name. After registration, the **Numbat admin** will activate your account.

User administration:

As an **admin** of the Numbat instance, navigate to the *'User administration'* section in the main menu. New users will appear with unverified email addresses. Verify their email

addresses and assign the appropriate *privileges* - *User* or *Admin*. The descriptions of these privileges are provided on the same page.

6.2   To set up the Openness extraction form in your Numbat workspace, select '*Edit extraction form*' and click on the '*Import an extraction form*' button. Ensure that the extraction form is in JSON format.
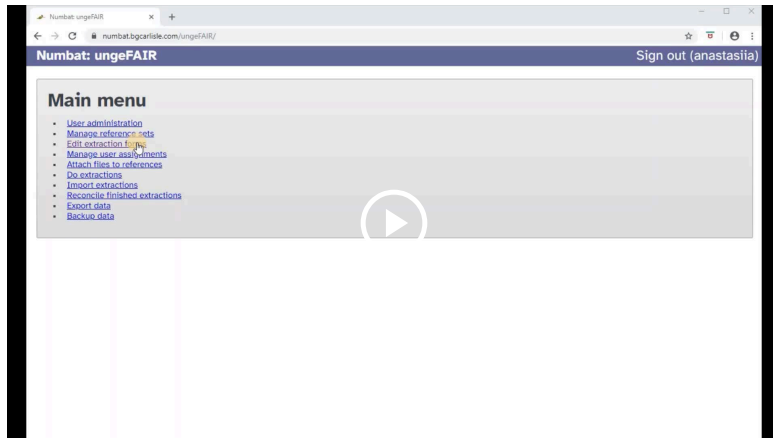
You may need to make additional adjustments, such as modifying conditions for when and how each question should appear.

6.3   **Prepare article list (ODDPub results) for Numbat**

Filter the output from ODDPub in the *is_open_data* column to include only **TRUE** statements. You can use a table calculation program like Excel or any text editor.

Save the filtered data in **.tsv** format (tab-delimited text file) as new input for Numbat. A text editor is more suitable than Excel for this task, as Excel spreadsheets are known to cause unexpected errors. For example, adding a small symbol or space can lead to incorrect file reading, even if the spreadsheet appears correct upon visual inspection. Additionally, copy and paste actions in Excel can result in lost or modified data. By using a text editor, you maintain control over every character.

7   **Load the article DOIs and detected statements into Numbat**

7.1   Select '**Manage reference sets**' section from the Numbat menu, then click on '**Add new reference set**'.
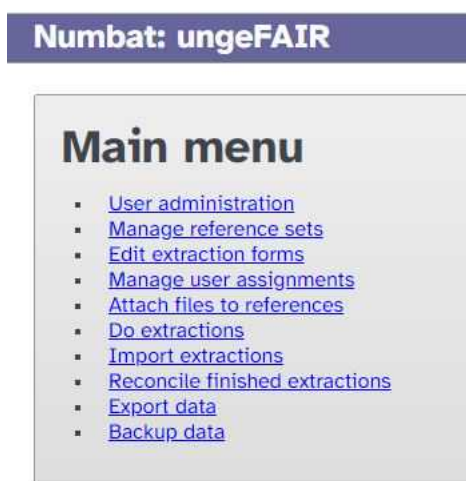
Fig. 7.1. Numbat main menu

Part of **SPRINGER NATURE**
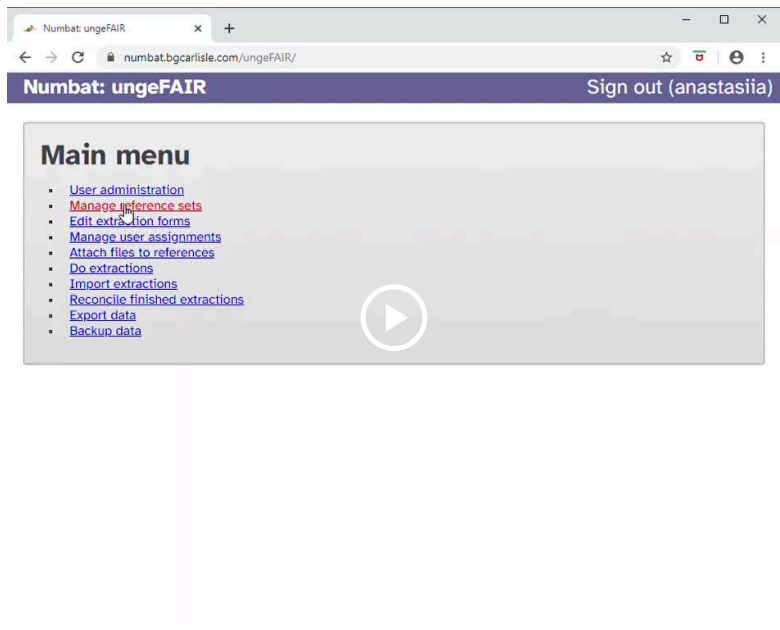
7.2   Select the relevant columns from the dropdown menu (e.g., *doi, is_open_data, open_data_statement, is_reuse*) and assign a name to the set (e. g. 'Publications of 2020').

Fig. 7.2. Process of adding a new reference set

7.3   Assign the new dataset to a user (or multiple users) via '**Manage extraction assignments'** in the **Main menu**:
1. Select either all records or choose specific ones from the list.
2. In the '*For the following form*' section, select the extraction form you want to use.
3. In the '*For the following user*' section, choose the user(s) to whom you want to assign the tasks.
4. Click '*Assign to user*' - a list will appear below, with successful assignments highlighted in green.

Fig. 7.3. Process of user assignment

Explore different assignment options, such as assigning articles that have already been screened by one rater or randomly selecting some articles.

Post-process extracted table in Numbat - Admin access

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

You can only export the data if more than two records have been checked.

Before downloading the results table, ensure that the answers have been reconciled if multiple users have checked the same data records. Otherwise, you may encounter duplicates.

1. Go to the '**Export data'** section → '**Export Openness extractions'** to download the results table.
2. Clean up the table:
- For Excel: go to the 'Data' tab, then 'Text in Columns' → celect 'Delimited' + tab stop / comma (depending on your data) + Standard → Finish. Save the file in your desired format.

Consider that the extraction dataset may contain several Open Data assessments per article. If you need an analysis at the article level, process the output accordingly. For example, when incentivizing data sharing at the article level, identify all publications in which at least one dataset was shared (openly or with restricted access).

> **Note**
>
> The output, in the form of a .csv file, can be downloaded any time from the Numbat server. This file contains all datarelated to criteria decisions as well as the final decision.

article	is_open_data	open_data_category	is_open_code	open_data_statements	open_code_statements	reference_to_data	comment1_reference_to_data	identifier	own_or_reuse_data	comment2_own_data	data_in_supplement	comment3_data_in_supplement	findability	comment4_findability	data_access	comment5_data_access	is_machine_readable_format	comment6_format	machine_readable_format_excel	machine_readable_format_csv	machine_readable_format_txt	machine_readable_format_spss	machine_readable_format_other_text_formats	machine_readable_format_video	machine_readable_format_audio	machine_readable_format_picture	machine_readable_format_fasta_fastq	machine_readable_format_raw	machine_readable_format_genetic_sequences	machine_readable_format_subject_specific_format	machine_readable_format_unsure	comment7_machine_readable_format	illustrative_files	comment8_illustrative_files	analytical_replication	comment9_analytical_replication	assessment	comment10_open_data_discussion
10.1038/s41467-020-16734-3	TRUE	general-purpose repository	TRUE	code availability all code used to analyze the dataset is openly available within	code availability all code used to analyze the dataset is openly available within	yes	NULL	supplement	own_open_data	NULL	no	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	no_open_data	or ya supplement availble e

		lead-dbs/-connectomesoftware (https://github.com/leaddbs/leaddbs).	lead-dbs/-connectomesoftware (https://github.com/leaddbs/leaddbs).																																			
10.1038/s41467-020-16929-8	TRUE	field-specificrepository	FALSE	proteomicsdatahavebeendepositedtoprideserv	NA	yes	NULL	http://proteomecentral.proteomexchange.org/c	own_open_data	NULL	yes	NULL	yes	NULL	yes	NULL	yes	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	1	NULL	NULL	NULL	NULL	yes	NULL	yes	NULL	open_data	NULL

				er under accession code pxd017341 [http://proteomecentral.proteomexchange.org/cgi/getdataset?				gi/GetDataset?ID=PXD017341																												
10.1186/s1291	TRUE	supplement	FALSE	additional file5.ka	NA	yes	NULL	TriNetX	unsure	NULL	unsure	NULL	unsure	NULL	restricted	not publicly availa	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

6-020-01851-z			plan-meier raw data										ble, the data are available from the authors upon reasonable request and with the permission of TriNetX.															

Tab. 8. Expected Numbat output table

Different issue handling in Numbat - Admin access

9   The extraction process up to this point has been linear. Steps 9.1 to 9.4 outline optional procedures. Sometimes, it may be necessary to correct errors (9.1), or add one or more datasets to the extraction list. A new article entry can be uploaded to the existing Numbat list (9.2). If an extractor is unsure about the Open Data status of a dataset, the extraction can be reassigned to another extractor (9.3). Assessments from two or more

extractors - whether for unclear cases or for quality assurance - can then be reconciled (9.4).

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

1. Go to the '**Do extractions**' section.
2. Locate the dataset you need to correct.
3. Open the extraction form as usual by clicking '*Extract*'.
4. Change the answer(s) as needed, then click on '*Completed*' if required.
5. If the extractions are finished and an export has already occured, delete the older version of the table from your storage and export the updated output via **'Export data'.**

> **Note**
>
> When correcting an extraction, all answers that depend logically on the affected answer will be **deleted** to prevent internal inconsistencies. Consequently, you will need to redo the extraction for all questions following the affected one.
>
> For example, if you change your response to the first question - whether there is a clear reference to the dataset in the article - from 'yes' to 'no', all subsequent questions (except the last one) will be skipped. The previous answers will be overwritten once you click 'Complete'.

9.2 How to **upload new article** to an existing reference set (table):

1. Create a new .tsv document in your text editor. This document should include the same column names as your existing reference set and contain all the references (DOIs) that must be added.
2. Go to '**Manage reference set'** → then '**Your set name XY'** → and **'new reference'**.
3. Verify that the columns in the updated table match the existing ones .
4. Assign the new records to one or more users via '**Manage user assignments**'.

9.3 How to **assign a questionable dataset** to another extractor:

At the end of every extraction, there is an '**Assign to'** option. Select the extractor(s) and the relevant extraction form (here: *Openness*), then finalize the assignment by clicking on '*Completed'*.

For our use case, we assigned datasets to other extractors whenever the Open Data status was marked 'unsure'. Conversely, we did not reassign datasets if the status was clear to the first extractor (i.e., 'yes' or 'no').

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

This function is essential when multiple extractors have worked on the same articles. If more than one extractor has completed the extraction, their answers can be compared and reconciled. Overlapping responses can be reconciled immediately, while differing responses may require further discussion or a decision based on the extractors' comments.

1. Go to the **'Reconcile finished extractions'** section.
2. The extractors who completed the extractions will be listed under the '*Extractors*' column.
3. Click on '**Reconcile extractions**':
   - compare the answers of two or more extractors, which will be displayed in adjacent columns.
   - a discussion among the extractors may be necessary to reach a clear conclusion, unless one extractor is designated as the 'master extractor' and undertakes the final decision. Pay attention to the commentaries, as they may include valuable insights for decision-making.
4. Save the selected answer in the final report by clicking **'Copy to final'.**
5. The final copy can still be edited and enriched with additional comments if none of the original answers is fully satisfactory.
6. Copy each sub-extraction individually into the final copy.

7. As in the article extraction, ensure to click '*Completed*' after finishing the reconciliation.

Numbat local installation - (System) Admin access

10 | **What is Numbat and how to install it**

> **Note**
>
> <u>Numbat</u> is a tool designed for extracting information from primary sources to assist in writing systematic reviews within an academic context and managing the resulting databases. It allows for assigning extraction tasks to multiple raters and reconciling their outcomes, meaning that it can compare and consolidate these into a joint assessment.
>
> Key Numbat functionalities:
>
> - Create an extraction form (questionnaire) that corresponds to your needs, such as a list of questions about datasets. This form is applied to each record.
> - Import information detected from sources like ODDPub into Numbat in the form of a .tsv table, which contains the records you wish to validate.
> - You can assign the dataset to multiple users if you want to compare extractions between raters, or assign it to a single responsible person.
> - The final dataset, with all answers to the extraction form, can be exported as a table.
>
> Numbat does not compute any statistics; its primary function is to collect information about records that would otherwise require fully manual entry without the aid of a semi-automated extraction form.
>
> Numbat is built on PHP and is free and open-source software under the GNU AGPL v3 license.

How to install Numbat (for Windows OS):

Requirements:
- Apache HTTP Server
- MySQL Database - provided by your company/organization
- PHP

> **Note**
>
> Installation on a **server** may require additional services and access rights; please contact your server's system administrator for assistance.

10.1 | **Clone the <u>Numbat repository</u> from GitHub locally**

You can use a GitHub account for this, or simply download the Numbat folder without logging in, as it is free open-source software.
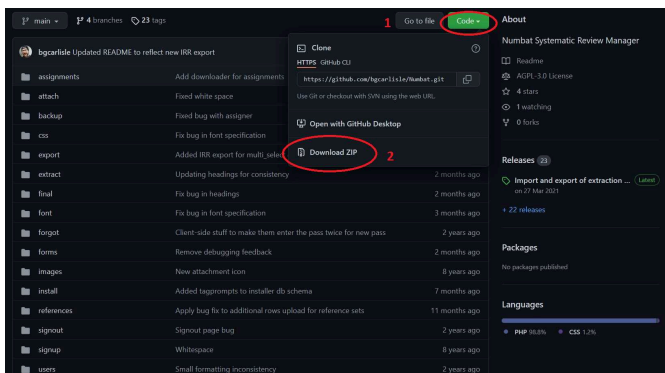

Fig. 10.1. Numbat repository clone from github

10.2 | **Install the XAMPP package**

It includes the Apache distribution and MySQL database.

Software	
XAMPP	NAME
Windows, Linux, OS X	OS
Apache Friends	DEVELOPER
https://www.apachefriends.org/de/index.html	SOURCE LINK

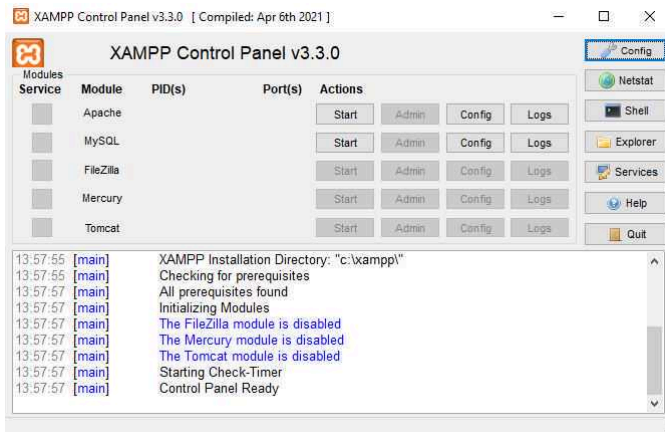This is what XAMPP looks like before any distributions are started:

Fig. 10.2. XAMPP program (inactive mode)

10.3 **Install Numbat in XAMPP**

- Copy the entire repository into the **'htdocs'** folder within XAMPP.
- **In XAMPP**, start the **Apache** module by clicking on '**Start**'.
- Click on '**Admin'** button next to Apache - you should see the installation instruction if they are not yet completed.

The following image shows a successful installation:
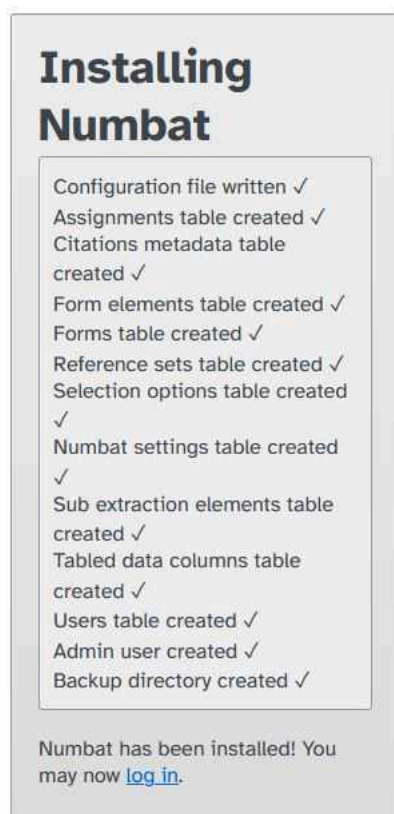
Fig. 10.3. Numbat installation via Apache admin

10.4 **MySQL setup** (probably provided by the company)

- **In XAMPP**: start the **MySQL** module by clicking on 'Start'.
- Click the '**Admin**' button next to MySQL to access the database.
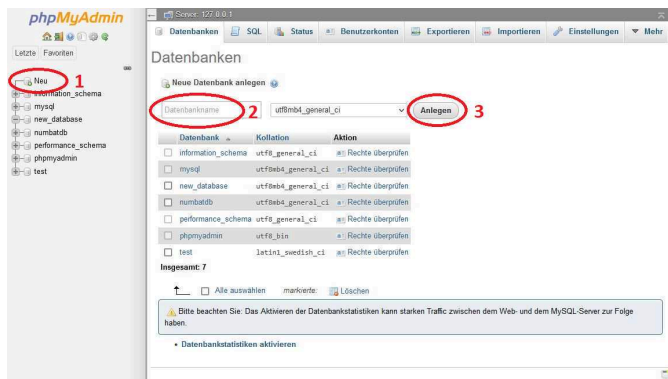- Create a new **MySQL** *database* for Numbat

Fig. 10.4.1. 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 Figure 10.4.3
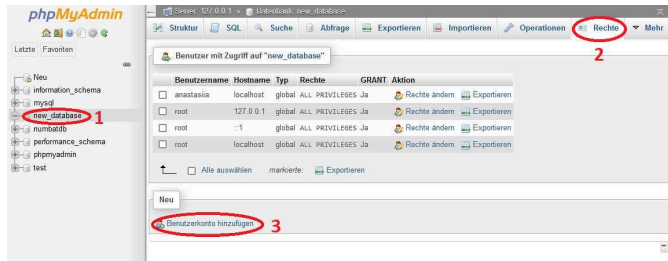
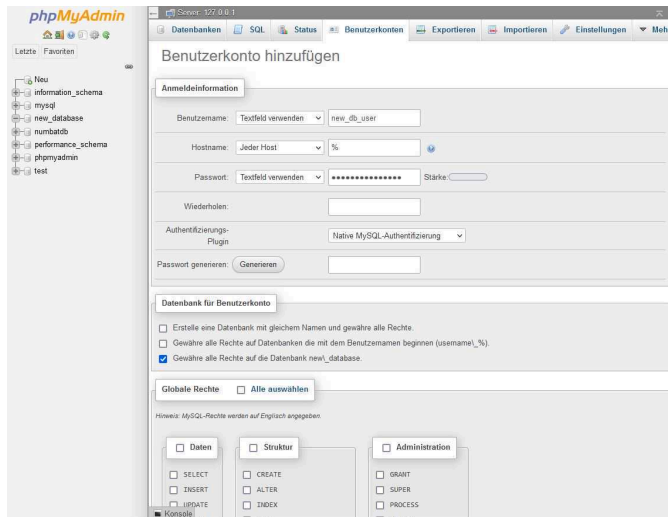Fig. 10.4.2. New admin user in MySQL

Fig. 10.4.3. Add user account in MySQL

10.5

- **In XAMPP**: use the **'Admin'** button next to Apache to complete the installation.
- In the window that opens, fill out the *database username*, *password, name*, and *host*, as well as the *URL name* - everything except the URL you created in step 10.4.

Citations

Step 1

Evgeny Bobrov, Nico Riedel, Miriam Kip. Operationalizing open and restricted-access data—Formulating verifiable criteria for the openness of data sets mentioned in biomedical research articles
10.1162/qss_a_00301

Step 3

Iarkaeva, A., Nachev, V., & Bobrov, E. . Workflow for detecting biomedical articles with underlying open and restricted-access datasets
10.31222/osf.io/z4bkf