Sep 09, 2024

A Protocol for Assessing Open Data Practices: Honours Students Can Lead the Way.

DOI
dx.doi.org/10.17504/protocols.io.kxygxyxmdl8j/v1

Haya Deeb[1], tomasz.zielinski[1], Andrew.Millar[1]

[1]Centre for Engineering Biology and School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK.

Andrew.Millar: Corresponding author

BioRDM

Haya Deeb
University of Edinburgh

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.kxygxyxmdl8j/v1

Protocol Citation: Haya Deeb, tomasz.zielinski, Andrew.Millar 2024. A Protocol for Assessing Open Data Practices: Honours Students Can Lead the Way.. **protocols.io https://dx.doi.org/10.17504/protocols.io.kxygxyxmdl8j/v1**

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

Created: September 09, 2024

Last Modified: September 09, 2024

Protocol Integer ID: 107180

Keywords: Data-Sharing, Open Data, Open Practice, Open Data Evaluation, Data Sharing Protocol

Funders Acknowledgement:
UKRI-MRC
Grant ID: funding award
MR/X009726/1

Abstract

Introduction

The culture of scientific research is increasingly recognising the importance of Open Data. Open Data practices involve making research data freely available for others to use, which promotes research integrity, collaboration, and resource efficiency. This protocol aims to assess the openness and FAIRness (Findable, Accessible, Interoperable, and Reusable) of data shared in bioscience research publications. By following this protocol, researchers can systematically evaluate the data-sharing practices in their institutions, thus contributing to a broader culture of open data.

This protocol documents the methodology used in the bioscience research field at the University of Edinburgh. The main purpose is to provide a comprehensive and replicable framework that can be adopted by other departments or research domains to evaluate and enhance their data-sharing practices. By sharing this methodology, we invite other institutions and departments to use this protocol to advance the culture of open data in their respective fields.

Objective and Goals

Objective

The primary objective of this protocol is to evaluate the data-sharing practices in research publications. This involves assessing the completeness, reusability, accessibility, and licensing of data associated with published papers; recording research-related variables such as data types; and publication-related variables such as data availability statements.

Goals

1. To provide a structured and consistent methodology for assessing open data practices in bioscience research.
2. To identify trends and challenges in data sharing across different research fields and over time.
3. To promote the adoption of open data practices by highlighting successful examples and areas needing improvement.
4. To encourage Honours students and interns to apply this protocol in their research projects, thus fostering a culture of open data from an early stage in their careers.
5. To encourage other institutions and departments to evaluate their data-sharing practices using this documented methodology, thus fostering a culture of open data across the biosciences.

Attachments

PDF
A Protocol for Asses...
555KB

XLSX
Data Extraction Shee...
12KB

Sampling Framework and Selection Process

1 **Research groups** within the biosciences at the University of Edinburgh were selected for study based on the educational interests of the undergraduate honours student researchers. Articles from each group were retrieved from the University's public Edinburgh Research Explorer website, which is based on a curated, institutional research information system.

2 **To be included** in the study, each research group needed to have a substantial publication record, with at least 10 journal articles published during the study timeline, between 2014 and 2023.

3 From each selected group, **journal articles were randomly chosen** for each year within the study period to ensure a representative sample. The focus was on original research papers that generated new datasets, **excluding other publication types** such as reviews of any type including systematic reviews, editorials, and commentaries.

4 We aimed to select **one publication per research group per year**. In cases where no papers were available for a specific year, additional papers from different years were selected to ensure that the number of publications for each research group remained consistent and close to each other.

Data Extraction

5 After selecting the sample from the research groups, the data extraction process was carried out in four main steps (Figure 1):

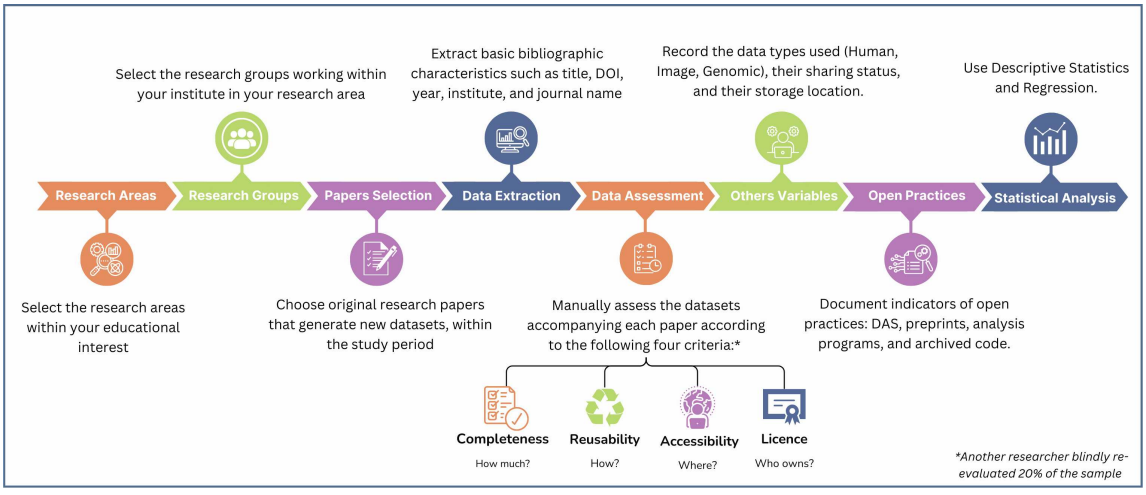

Figure 1: Overview of the Methodological Steps of this protocol

5.1 **Step 1: Main Characteristics of the Paper:**
Extract basic bibliographic details such as title, year of publication, institute or research group, research field, and journal name.

5.2 **Step 2: Dataset Evaluation Based on Four Criteria:**
Manually assess the datasets accompanying each paper according to the following criteria: Completeness, Reusability, Accessibility, and Licence.

5.3 **Step 3: Key Characteristics of the Research Data:**
For each paper, additional information about the data types was extracted. This included whether specific types of data (Human, Image, Genomic) were used and if they were shared when used. Furthermore, the location where the data was shared and stored, such as in a repository or as supplementary material, was documented.

5.4 **Step 4: Other Open Indicators:**
Document additional indicators of open practices, such as the presence of data availability statements, preprints, the use of analysis programs, and whether the code used in the research is archived.

6 To facilitate this process, we have provided a template along with a README file (attached to this protocol), which can be used to extract the variables.

We highly recommend using the same template to ensure that your work can be effortlessly combined with ours. Additionally, any new variables that could benefit your research domain and enhance open data practices can be added separately. This approach maximises the reproducibility and utility of this methodology across different studies and research domains.

Data Extraction Sheet & README Fil… 12KB

Step 1: Main Characteristics of the Paper:

7 For each selected paper, the following key characteristics were extracted:

1. **Title:** The exact title of the paper was copied and pasted as it appears in the publication. This ensures
accuracy and consistency in identifying the document.
2. **DOI (Digital Object Identifier):** The DOI link of the paper was recorded. The DOI is a Persistent Identifier (PID), a unique alphanumeric string that provides a permanent link to the document administered by an external organisation, distinct from a direct URL link, and uniquely identifies the document for easy retrieval.
3. **Year of Publication:** The year of publication was noted down as a four-digit number. This helps in tracking the temporal aspect of data sharing practices.

4. **Research Institute:** The associated research institute or department for each selected paper was determined. This information aligns with the predefined research interest of the project. Depending on the study's focus, this could be either the specific institute being focused on or the name of the research group leader prominent in that field, as indicated in the author affiliations. For example, at the University of Edinburgh, the institutes covered included the School of Biological Sciences and the College of Medicine.
5. **Principal Investigator (PI):** The surname of the Principal Investigator (PI) or the leader of the research group was recorded. While this detail may not be crucial for the publications themselves, it was useful for internal meetings, evaluations, and discussions.
6. **Research Areas:** The primary research area covered in the paper was specified. This might be consistent across the entire project depending on its focus. In our study, the research areas were: Non-Communicable Diseases (NCDs), Infectious Diseases and their Treatments (InfD), Microbial Biotechnology (Biotech), Stem Cells and Regeneration Medicine (SRM)
7. **Journal Name:** The name of the journal in which the paper was published was provided. This might be used to analyse the journal's open-access policies in future.

Step 2: Dataset Evaluation Based on Four Criteria:

8 The datasets accompanying each paper were manually assessed according to four key criteria: Completeness, Reusability, Accessibility, and Licence. This evaluation aimed to determine the quality and openness of the shared data.

The assessment of shared data was conducted regardless of where the sharing was mentioned, whether in a specific data availability statement, within the methods section, or elsewhere in the article. Similarly, the data's location—whether stored in repositories or included as supplementary material—was also considered. These criteria were designed to evaluate the overall openness and FAIRness of data sharing practices.

Before starting the scoring process, it is important to note that **Completeness** is evaluated based on all the data provided in the article. In contrast, **Reusability, Accessibility, and Licence** are assessed only on the most comprehensively shared dataset. For example, if a paper includes multiple datasets but only shares a subset, the Completeness score reflects this partial sharing, potentially resulting in a lower score (e.g., 3 or 2). However, the assessments for Reusability, Accessibility, and Licence focus on the dataset that has been shared in the most complete and usable form (the best shared dataset).

All criteria are scored on a scale from 1 to 4, with 1 being the lowest score and 4 being the best score (detailed in Table 1).

8.1 **Completeness:**

Completeness reflects whether all the datasets needed to reproduce the results were shared.

- Score 1: Indicates that no data has been shared. In such cases, a score of 1 ("Not Scored") is automatically assigned to the remaining criteria (Reusability, Accessibility, and Licence).
- Score 4: Represents high completeness, where datasets are fully accessible, shared, and can be downloaded in any format. If not all the datasets are shared, the score is adjusted to 3 or 2, depending on the extent of data availability.

The Completeness score thus indicates how thoroughly the research data has been shared, relative to the total data reported or analysed in the publication. Assessing the total, i.e. what data should be shared, requires some understanding of the research methods. In our case, final-year undergraduate students assessed publications from the research areas that they had studied.

8.2 **Reusability:**

The Reusability criterion focuses on the file type of the shared dataset to ensure that the data is in a user friendly format. This includes checking whether the data is stored in non-proprietary, human- and machine-readable formats, such as .xlsx or .csv, which facilitate data aggregation and can be processed with both free and proprietary software. Table 2 provides further details on file formats, indicating whether they are human- or machine-readable and whether they are non-proprietary. Additionally, the presence of metadata or a ReadMe file is essential, as these provide necessary context and instructions for understanding and using the data.

8.3 **Accessibility:**

The Accessibility criterion evaluates where and how the data is shared. This includes assessing the ease of access to the data and ensuring that it is readily available for use without unnecessary barriers. Key factors include whether the data is stored in an online public repository and whether it has a Persistent Identifier (PID) such as a DOI, or a unique identifier that is internal to the repository such as a GenBank GI, that ensures the data can be easily located and accessed by other researchers. The goal is to ensure that the data is as open and accessible as possible, facilitating further research and reuse.

8.4 **Licence:**

The Licence criterion assesses the licensing conditions under which the dataset is shared, which may differ from the licence of the article itself. The licence should be evaluated based on the criteria of the public platform where the data is stored, or as specified when shared. If the data is included in supplementary files, it is assumed that the dataset follows the same licence as the article, whether it is open access or restricted. Some journals use a different licence for their supplements as a standard policy but the policy might not be linked from the article and can take time to discover.

Be cautious when determining the journal's open access status, as institutional access (e.g., through a university account) may not accurately reflect the journal's true open access nature. Always confirm the licensing details provided in the article, either at the beginning of the text or within the PDF.

8.5

Table 1: Scoring Criteria used for the Assessment of Articles		
Data Completeness		
Score	Description	Criteria
4	Exemplary	All the data necessary to reproduce the analyses and results (in practice) are present within the article (e.g. as supplementary information, figures) or archived in external repositories. Both raw and processed datasets from all methods utilised and mentioned in the study are provided.
3	Good	Most of the data necessary to reproduce the analyses and results (in practice) are present within the article (e.g. as supplementary information, figures) or are archived in external repositories. Processed data from all methods utilised and mentioned in the article are provided, lacking only a small amount of raw datasets.
2	Average	Main analyses in the paper cannot be redone because essential datasets are missing AND/OR only summary statistics (e.g. means, standard deviation) obtained from methods utilised and mentioned in the article are archived, no raw data provided.
1	Poor	Neither processed nor raw data are archived/present in the article OR the incorrect data are archived.
Reusability		
Score	Description	Criteria (For best dataset)
4	Exemplary	Good formats and metadata. Data is archived in a (1) non-proprietary, (2) human- and machine-readable file format that facilitates data aggregation and can be processed with both free and proprietary software (e.g., csv, text); other formats were allowed as de facto community standards (see Table 2). And (3) Highly informative metadata (such that column headings, abbreviations, and units can be understood in isolation from the original paper).
3	Good	Data is archived in (1) a non-proprietary OR (2) a human- and machine-readable file format that facilitates data aggregation and can be processed with both free and proprietary software. OR (3) Metadata must at least be sufficiently informative to be understood when combined with the paper.
2	Average	Poor formats and metadata. Data is archived in (1) a proprietary OR (2) human- but not machine-readable file format (e.g., pdf, jpeg). AND (3) Metadata is not sufficiently informative when combined with the paper.
1	Poor	Not Scored
Accessibility		
Score	Description	Criteria (For best dataset)
4	Exemplary	Data is accessible, has a Persistent Identifier (PID) assigned or a unique identifier from a high-tier repository (such as a GenBank genInfo number), and is stored in an online public repository (e.g., Figshare).
3	Good	Data is accessible, has either a PID or a unique identifier from a high-tier repository OR is stored in an online public repository.
2	Average	Data is accessible but does not have a PID or unique identifier and is not stored in an online public repository (e.g. data shared in the supplementary information section of the article).
1	Poor	Not Scored
Licence		
Score	Description	Criteria (For best dataset)
4	Exemplary	Data has a permissive licence (e.g. CC0, CC-BY) and code has an Open Software licence (if applicable).

3	Good	Licence(s) are present but not all are permissive. Data has a restrictive licence (e.g. CC-BY-SA/NC) and/or code has a Closed Software licence (if applicable).
2	Average	No explicit licence is provided for data or software.
1	Poor	Not Scored

Table 1: Scoring Criteria used for the Assessment of the Articles

9

File Extension	Non-Proprietary	Human-readable	Machine-readable
.ab1	1	0	1
.avi	0	1	NA
.bam	1	0	1
.csv	1	1	1
.doc	1	1	0
.docx	1	1	0
.fastq	1	1	1
.gif	0	1	NA
.jpg	1	1	NA
.maf	1	1	1
.mov	0	1	NA
.mp4	0	1	NA
.pdf	0	1	0
.raw*	0	1	1
.rtf	0	1	0
.sas	0	1	1
.sav	0	1	1
.tif	1	1	NA
.txt	1	1	1
.vcb	1	1	1
.wav	0	1	0
.xls	1	1	1
.xlsx	1	1	1
.xml	1	1	1

Table 2: Characteristics of data file formats. In this table, "0" is no, and "1" is yes. NA was regarded as "yes" when scoring for "Data Reusability".

(*.raw = A camera raw image file contains unprocessed or minimally processed data from the image sensor of either a digital camera, a motion picture film scanner, or other image scanner)

This table was adapted from Roche et al (Roche et al., 2015).

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biology, 13(11), e1002295. https://doi.org/10.1371/journal.pbio.1002295

Table 2: Characteristic of Data File Formats.

Step 3: Key Characteristics of the Data

10 For each selected paper, additional information about the data used in the research was extracted, as follows:

1. **Image Data Sharing:** If no image-related analysis (including video analysis) was applicable or conducted, 'NA' was entered. A score of '0' was recorded if images were analysed but no related raw data were shared. A score of '1' was assigned if any raw image data (including videos) were shared, regardless of completeness.
2. **Genomic Data Sharing**: 'NA' was noted if genomic studies (such as sequencing, gene expression analyses, or microarray studies) were not performed. A score of '0' was given if any genomic data were produced but none were shared. A score of '1' indicated that some genomic raw data were shared, even if not all data were provided. Note that certain DNA sequences, such as PCR primers or Taqman assay probes, were not scored as genomic data.
3. **Proteomic Data Sharing:** If no proteomic investigations (like mass spectrometry analyses) were conducted, 'NA' was entered. A score of '0' was given if proteomic data were generated but not shared in any form. A score of '1' was assigned for any level of raw proteomic data sharing, irrespective of completeness.
4. **Human Patient Samples or Data:** If the research did not involve human patient samples or data, 'NA' was recorded. A score of '0' was given if human data were used but not shared. A score of '1' was assigned if any raw data from human patients were shared, even partially.
5. **Data Storage Location:** The platform or location where the data was shared was documented as a storage variable. This was categorised into:
   - Repository: If the data was shared exclusively in one or more public repositories, such as NCBI, Zenodo, etc.
   - Supplementary Material: If the data was shared only in the supplementary material of the publication.
   - Both: If datasets were shared both in a repository and as supplementary material.
   - NA: If no data was shared, this was recorded as 'NA'.

These details were recorded in the provided Excel sheet for the project's data extraction.

Step 4: Other Open Indicators

11 In addition to the evaluation of data, other indicators of open practices in the publication process were documented:

1. **Analysis Programme:** This variable records whether the research paper specifies the statistical or other analysis software used, such as SPSS, GraphPad, R, SAS, and Stata. The type of software is not scored, nor whether it is a standard product or a bespoke programme. This information is typically found in the 'Methods' section of the paper, particularly under statistical analysis. This is coded as '0' for no and '1' for yes.
2. **Code Archived**: This variable indicates whether the scripts or code used for analysis are archived in a public repository or available as supplementary material. If no code was used, this is marked as 'NA.' If code was used but not archived, it is marked as '0,' and if it is archived, it is marked as '1.' Code availability may be detailed in sections titled 'Code Availability' or 'Data and Code Availability,' but could also be found in other parts of the paper, including 'Statistical Analysis,' 'Author Notes,' or 'Acknowledgements.' External repositories like Zenodo should also be checked for code storage.
3. **DAS (Data Availability Statement):** The presence of an explicit Data Availability Statement (DAS) is checked, which may be found in sections titled 'Data Availability Statement,' 'Data Availability,' 'Materials and Methods Availability', etc. If there is no DAS, the variable is marked as 'NA.' If a DAS is present and indicates that data is available upon request, it is marked as '0.' If the DAS clearly states that all data is available on specific platforms or within the paper, it is marked as '1.'
4. **Corresponding Author:** This variable determines if the group leader of the selected research group is listed as a corresponding author on the paper. It is marked as '0' for no and '1' for yes.
5. **Preprints:** The presence of a preprint version of the article prior to journal publication is identified. Scholarly databases (such as PubMed and Google Scholar) and preprint servers (such as BioRxiv and MedRxiv) are used to determine whether earlier versions of the paper were uploaded before peer review. Note that titles and authorship lists can vary among article versions. This variable is marked as '0' for no and '1' for yes. These indicators provide a broader view of the open practices associated with each paper, complementing the dataset evaluation.

These indicators provide a broader view of the open practices associated with each paper, complementing the dataset evaluation.

Data Extraction and Quality Control Assessment

12 The data extracted from the selected papers, including both scoring assessments and supplementary variables, were systematically catalogued in an Excel spreadsheet. A template of this spreadsheet has been shared with the protocol to facilitate consistency in data collection. Multiple students can work in parallel, each on a different research area, with shared supervision meetings to ensure consistent scoring.

To ensure the reliability and accuracy of the assessments, a quality control process was implemented. Specifically, 15-20% of the papers from each evaluator's dataset were randomly

selected for reassessment. This selection was supplemented to ensure coverage across the range of scoring values and publication years.

The reassessment was conducted by a different researcher who was blind to the original scores. Students also partially re-scored each others' work to test reproducibility, and this could increase to full double scoring. It is highly recommended that the comparison between the initial and secondary evaluations achieve an agreement of no less than 85%. Based on this level of agreement, the initial assessments can be considered reliable and retained for the final analysis.

Statistical Analysis

13 Descriptive statistics were presented as frequencies and percentages (n(%)) to summarise the data. The sample was split by research areas in this project, and it is recommended that future studies segment their samples in ways that best address their specific research questions. Stacked bar charts were used to visualise the overall scores for the four criteria, as well as changes in these scores over the study period.

Ordinal regression models were employed to assess changes in the four scoring criteria over time, with year as the independent variable and the score as the dependent variable. Additionally, ordinal regression was used to evaluate the influence of the research area on the scoring criteria. The final ordinal regression model assessed the impact of other sharing variables, such as Data Availability Statements (DAS) and preprint status, on each scoring criterion, with DAS and preprint status as independent variables and the scoring criteria as the dependent variables.

To account for potential intra-group similarities in data-sharing practices, random effects for research groups were included in the models, acknowledging the possibility of consistent behaviours within groups. Statistical significance was set at $p < 0.05$, and results were expressed as odds ratios with corresponding 95% confidence intervals. Details of the assumption check for the models, along with the complete analytical methodology, can be found on our project's GitHub repository: https://github.com/BioRDM/InsightsOfOpenPracticesInBiosciences.

All statistical analyses were performed using R and RStudio Software (Version 4.2.2).

Acknowledgements

14 We extend our sincere gratitude to Professor Simon N. Wood, Chair of Computational Statistics at the School of Mathematics, and his PhD student, Antoni Sieminski, for their invaluable support and expert consultation provided through the statistics drop-in clinics.
We thank Dr. Megan A.M. Kutzer for statistical input, the Honours students who participated in protocol development and/or testing, Suzanna Creasey, Diego Lucini de Ugarte, George

Strevens, Trisha Usman, and Hwee Yun Wong; and UKRI-MRC for funding award MR/X009726/1.

Standpoint

15 The authors' training and expertise comprise clinical medicine, mental health practice and data science (HD), research software engineering (TZ), experimental biology (AJM), and research data management (HD, TZ, AJM). The Centre for Engineering Biology (formerly SynthSys, and the Centre for Systems Biology at Edinburgh) is an interdisciplinary research centre hosted by the School of Biological Sciences.
AJM was its founding, Director. The University of Edinburgh is one of the largest research-intensive universities in the United Kingdom. The authors work in the University's Biological Research Data Management team, BioRDM. Undergraduate Biology students in their final year at the University, known as 'Honours students', conduct a 3-4 month research project. This protocol was initially developed with and/or applied by six Honours students, whose project research was supervised by AJM.