Data Guidelines

1. Overview

AAS Open Research requires that the source data underlying the results are made available as soon as an article is published. This page provides information about data you need to include, where your data can be stored, and how your data should be presented.

A large number of journals and publishers have confirmed that they welcome research articles reporting analysis and conclusions that are based on previously published datasets: They do not consider the publication of a dataset with a DOI and associated protocol information as a 'prior publication' that would preclude subsequent publication of new results obtained from such a dataset.

2. Requirements

All articles should include the submission of the data underlying the results, together with details of any software used to process results. It is essential that others can see the raw data to be able to replicate your study and analysis of the data, as well as in some circumstances, reuse it. Furthermore, publishing your data will show clearly that you did the work first. Others that then reuse your data for their own studies will be required to cite your data (which can be cited separately from your article if appropriate). We ask authors to deposit their data with approved data repositories (see list below), normally under the CC0 licence which facilitates data reuse. Failure to provide such data for publication without good justification is likely to result in your article being rejected.

AAS Open Research endorses the FAIR Data Principles to make data Findable, Accessible, Interoperable and Re-usable. Authors may find it useful to consult FAIRSharing for details of additional data standards specific to their field of interest.

Exceptions: We recognise that there may be cases where openly sharing data may not be feasible (due to ethical or confidentiality considerations), or because the data have been obtained from a third party and access restrictions apply (see data policy for details). If you think that you cannot provide the source data, please let the editorial team know, so we can advise further.

3. Data Preparation Guidelines

When depositing data involving human participants, authors must ensure that all datasets have been de-identified in accordance with the Safe Harbor method before submission.

Please ensure that all files are labelled clearly so readers will understand the contents of, and difference between, the files. For each file/group, we suggest you provide:

- A single short title describing the content of the files;
- A more detailed legend describing each dataset, so it is clear that the files are distinct and downloadable (including the explanation of any acronyms used in the dataset).

In the manuscript, please provide a brief summary of the deposited datasets under a heading “Data Availability”.

3.1 Spreadsheet data

To increase the accessibility and reusability of spreadsheet data (i.e. large tables or raw data), they should adhere to the following best practices:

DO

- Give each column a descriptive heading.
- Use a single header row.
- Ensure you have used the first cell, i.e. A1.
- Include a title and a legend to describe each spreadsheet.
- Save each data file with a name that appropriately reflects the content of that file.
- Submit each table that is part of the dataset as a separate file.
- Submit each worksheet as a separate file.

DO NOT

- Embed charts, comments or tables within a spreadsheet.
- Use color coding (machine-based data mining cannot interpret this).
- Include special (i.e. non alphanumeric) characters within the spreadsheet, including commas.
- Use merged cells.
- Submit multiple worksheets within a spreadsheet (such as those used in Microsoft Excel), as these are not supported by CSV and TAB formats.

Spreadsheets should be submitted in CSV or TAB format; EXCEPT if the spreadsheet contains variable labels, code labels, or defined missing values, as these should be submitted in SAV, SAS or POR format, with the variable defined in English.

3.2 Software source code

All articles should include details of any software that is required to view the datasets described or to replicate the analysis. For all software used, please state the version used, details of where the software can be accessed, and any variable parameters that could impact the outcome of the results.

Where software has been coded by the authors of the paper, the source code should be made available. If there are ethical or privacy considerations as to why the source code may not be made available, please contact the editorial team.

Authors may find it useful to consult FAIRsharing.org for details of additional data standards specific to their field of interest.

4. Data Hosting

Data must be hosted by a stable and recognised open repository. Using such a repository ensures that your dataset continues to be available in a useable form in the future.

If there is a suitable subject-specific repository for your data, please deposit them there and include the Accession Number(s) or other Identifiers and database details in your manuscript. For some data types, such as genetic sequences and protein structures, it is essential that the data are deposited in GenBank and Protein Data Bank, respectively. For X-ray crystal structures, please also submit your validation reports.

If there is no obvious subject-specific repository for your data, we will be happy to advise on suitable options. We will need to know which file types you have and the approximate total size of your datasets. We will then discuss with you the best way to make your data available.

4.1 Non-exhaustive list of AAS Open Research-approved repositories

Below is a list of repositories that have already been approved for hosting data alongside a AAS Open Research article.

Please include the name of the repository used at the end of the manuscript, along with details indicated in the ‘What to include in the data availability section of your article’ column in the table below.

If you are an author who wishes to use a repository not already on this list, or you manage a repository that you would like included on the list, please contact us.

General data, research materials and supporting documents

DATA TYPE	WHERE TO SUBMIT*	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Any, but especially humanities/social science data and data in SAV and POR formats	Dataverse	Title, DOI
Any	Figshare	Title, DOI
Any, but especially deposits with mixed data, materials and documents	Open Science Framework†	Title, DOI
Any, but especially deposits with mixed data and code	Zenodo	Title, DOI

* Please note that many repositories have a limit on the size (usually 2 or 5 GB) of single file uploads and charge for larger data files.
† Deposits must be made public.

Humanities and social science data

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Any	DANS-EASY*	Title, DOI

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Any, but reserved for ISCPR member institutions	Open ICPSR	Title, DOI
Any	UK Data Archive*	Title, DOI
Social and economic data	UK Data Service	Title, DOI

* Deposits must be open access.

3D-printable models

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
All 3D-printable models (including molecular, cellular, medical/anatomical and labware models)	NIH 3D Print Exchange	Title, model ID, URL

Environmental and ecological data

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Complex environmental and ecological data	The Knowledge Network for Biocomplexity*	Title, DOI

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Environmental data collected by NERC-funded researchers	NERC data centres	Data centre name, title and DOI
Geospatial	PANGAEA	Title, DOI

* Data entries must be made public.

Sequence and omics data

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Expression and sequence data (including Nucleotide / protein sequence, microarray, SNP / SNV, GWAS, phenotype or sequence-based reagent data) Systems and chemical biology data (including chemical entities, chemical reactions, computational models, metabolic profiles, or molecular interactions)	Any appropriate NCBI- or EBI- based repository*	Accession number(s). For SNP / SNV data please provide HGVS name(s), local ID(s) and rs / ss number(s)

* Some higher-level repositories, such as BioProject, provide access to data deposited in various archival databases. In these cases, please cite the accession numbers that are assigned to the data submissions by the archival databases in addition to the higher-level identifier.

Health data (restricted access to protect anonymity of participants)

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Addiction and HIV data	National Addiction & HIV Data Archive Program	Title, DOI
Cancer imaging	Cancer Imaging Archive	Title, DOI

Macromolecule structures

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
3D protein structures	Protein Data Bank	PDB number
Crystallography*	Crystallography Open Database	COD ID
X-ray images	Coherent X-ray Imaging Data Bank	Title, DOI

* X-ray crystallography validation reports should be submitted (as a PDF) directly to AAS Open Research via the submission system.

Neuroimaging data

DATA TYPE	WHERE TO SUBMIT	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Raw fMRI datasets	OpenfMRI	Title and accession number(s)
MRI and PET unthresholded statistical maps	NeuroVault*	Title and URL (which includes a unique data ID)

* Please note that authors will still be expected to deposit their raw neuroimaging data in an appropriate repository. Also, once submitted, administrative powers will be transferred to AAS Open Research. This is necessary to ensure stability of the dataset; this transfer does not affect the CC0 licence assigned to all NeuroVault submissions.

Software & source code

DATA TYPE	WHERE TO SUBMIT*	WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE
Latest source code	GitHub, BitBucket, SourceForge or Google Code	URL
Archived source code	Zenodo	Title, DOI and licence* used
Software	Authors may host software where they wish, though it is strongly recommended to use a stable URL	URL

* An open licence must be assigned and we strongly advise authors to use an OSI-approved license.

AAS Open Research can accept papers with the underlying data being hosted by an approved institutional data repository. For other institutional repositories, please contact us.

4.2 Data repository requirements

In order to host data linked to an AAS Open Research article, a repository must be actively managed. Repositories must:

4.2.1 *Enable access to the dataset*

- Access to the data should normally be completely open, unless there are genuine concerns over security/privacy of the data. Information should be provided about who can access the data, terms and conditions of access, and a clear point of contact.
- The repository must have a policy for data that do require additional protection. This includes appropriate access for peer reviewers, as required as part of the data peer-review process. (In the context of data, peer reviewers are experienced researchers who produce or use data in the same field as the data being published.)

4.2.2 *Ensure dataset persistence*

- The repository must have a clear and public assertion of responsibility to preserve the data and provide access to the data over the long term.
- There must be an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate, or the governing or funding institution substantially changes its scope.
- The curators must develop and implement suitable quality control measures to ensure the metadata are correct and the data themselves are maintained and curated to avoid degradation. User feedback can and should be used to strengthen and correct the metadata as needed.
- Globally unique persistent IDs must be assigned to the published datasets and a repository-managed URI associated with each of those IDs must be maintained. These URIs should also be associated with versions of the datasets.
- Permanent IDs for the dataset must resolve to a publicly accessible landing page which must:
    a. be open and human readable (and it would be preferred that they should also be provided in a format which is machine readable)
    b. describe the data object and include appropriate metadata and the permanent identifier (used to identify the page in the first place)
    c. be maintained, even if the data have been retracted.

4.2.3 *Ensure dataset stability*

- Stability means that the exact same version of the dataset that was cited can be returned to when the citation is resolved.
- If dataset versioning is supported, new versions should be permanently identified and linked from the original, published dataset landing page, without overwriting the original version linked from the article. The database should provide time-stamped versions of archival data.

4.2.4 *Enable searching and retrieval of datasets*

- The repository must allow users to easily determine whether a dataset has been peer reviewed or been subject to an equivalent level of scientific quality assurance.
- It must provide appropriate metadata about the dataset in human readable form on the landing page, and when possible standardized machine readable formats e.g. DataCite metadata schema http://schema.datacite.org
- Access must be given to allow metadata for the datasets to be searched and retrieved through interfaces designed for both humans and computers.

We recommend that repositories also collect information about repository statistics:

- Publish statistics on the level of access to any deposited item that is publicly accessible, to contribute to metrics of the item's publication impact.
- Publish information that enables journals and depositors to assess the take-up in the community it aims to serve, e.g. about any operational agreement with a well-established journal, learned society or equivalent body.