Guidance for Data Providers: Supporting Documentation

This guidance document outlines what supporting information is required when submitting a dataset for deposit with the EIDC.

Why is supporting documentation required?

Supporting documentation is essential as it helps others understand a dataset and supports its potential re-use. This information should therefore be as detailed as possible.

What information is required?

As specified in the NERC Data Policy guidance notes, supporting documentation should provide information on the following areas:

Experimental Design/Sampling Regime

Metadata should be provided which details the experimental design and/or sampling regime, where applicable. This should include information on:

- The feature(s) of interest - including feature type, feature name, relevant geographical information and grid or reference system used e.g. location, aspect, elevation, surface area, volume, etc.

- The treatments applied - including details of how treatments will be applied/created/managed or verified, where relevant.

- Replication - details of any sample/observation replication, including explanations for any missing samples/observations.

- Controls - information on any control methods employed.

- The periodicity of sampling - details of the time period covered by the dataset, date and frequency of sample collection/observation recording and any reasons for missed sampling/observations.

- The number of samples/observations - the total overall number of samples/observations collected, if not already included as part of metadata relating to treatments and replication.

Collection/Generation/Transformation Methods

Information should be provided covering methods used for collection of samples/observations. This should include details of sampling/observation locations, relevant techniques employed for physical collection of samples or measurement of observations, sample storage/treatment or recording of observations, and also the details of the parties responsible for sample collection/observation measurement. Standard Operating Procedures (SOPs) for specific techniques and/or references for methods used should be included, if available. Date of analysis of samples should be included if different from date of sampling.

Alternatively, where data values are derived/ generated/ transformed, then details of how this is achieved should be provided. For model output, this should include information relating to the key points of the theory forming the basis of the model. The type of model used (e.g. ordinary differential equations, partial differential equations, compartment model) should be documented and any relevant technical information (e.g. operating system(s) and programming language) and/or mathematical information (e.g. input and output) used to generate the output also documented.

Fieldwork and/or Laboratory Instrumentation

Information should be supplied on instruments/machines used for collection/analysis of samples/observations where relevant. This should include the type, make, model and serial number of each particular instrument/machine, where known.

Calibration Steps and Values

Details of the steps taken to calibrate any instruments/machines used, including use of any blanks, and the values used for calibration should be provided.

Nature and Units of Recorded Values

Information should be provided describing the nature of the recorded values contained and the units used sufficient to unambiguously define what has been measured and recorded in the dataset. Details should include description of the parameters, determinands, variables, valid range of values, lowest level of detection, units etc.

Analytical Methods

Full descriptions of any analytical methods used to generate the data values contained in the dataset should be included. These should detail any reagents and the specific conditions required for each analysis, and provide sufficient detail to enable replication of the methods used for analysis if desired.

Quality Control

Any quality control measures undertaken to ensure the quality of the data values included in the dataset should be detailed. E.g. methods of quality control, explanation of quality codes, factors affecting the data.

Details of data structure

Details of the structure of the dataset should also be provided, covering the order in which variables appear within the dataset. For example: 'This dataset comprised 6 MS Excel spreadsheets, the spreadsheets were entitled xxx, xxx xxx etc., the first spreadsheet had 5 columns labelled xx, xxx etc.'. Or, 'This dataset comprised 4 tables in an Oracle database, the columns in table 1 were xxx, xxx, xxx etc and the tables were related in the following way…'.

Miscellaneous

Any additional information necessary to expand on that given in the discovery metadata record.

The documents containing this information may already exist in one or several places/formats e.g. in a technical report, in a spreadsheet alongside the data, on a project website or wiki, or the information may be integral to the data itself (e.g. netCDF format).

How should I supply the information?

In order to meet funders' expectations and the needs of future research, supporting information must be openly accessible in perpetuity.

In order to guarantee this the EIDC requires supporting documentation to be submitted prior to or at the same time as submission of data. Unfortunately it is not acceptable to link to pages/documents on non-EIDC websites, or to include documents with hyperlinks to external websites. This is because we are unable to guarantee that those websites/pages will exist in perpetuity.

Formats

Our preferred formats are RTF, HTML, .txt or .csv.

If the information is stored in a proprietary format such as Microsoft Word it should be converted to one of the preferred formats.

Stand-alone Documents

Supporting documentation should be supplied separately from data. This is to ensure detailed metadata are available without the need to download the data, and in perpetuity. Making the data and metadata available separately ensures the EIDC are able to securely store the data and ensure it remains unchanged whilst being able to continually improve the quality and usefulness of the contextual metadata.

If the information is integral to the data itself (e.g. netCDF) it should be copied into a document in a preferred format.

Filenames

Names of files supplied must not include any spaces or non-standard characters.

Additional notes:

Often depositors ask if they can provide their scientific papers published through journals as supporting documentation for the dataset. Unfortunately these do not meet the requirements for one or all of the following reasons:

- Unsustainable format – journal papers are often supplied in PDF or similar which is not a preferred format.

- Lacking dataset description - the dataset is not described adequately in a publication describing a scientific research outcome e.g. it is unlikely that the column names are described.

- Open web-access cannot be guaranteed in perpetuity – the EIDC cannot guarantee permanent, open access to external websites/pages in perpetuity as they have no control over the pages e.g. if the server/page address changes, if the server is faulty/ taken down etc.

Where the published paper is in a long-term format and describes the dataset adequately it may seem the obvious approach would be for a copy of the actual journal paper to be lodged. However, CEH would need to be absolutely sure that it was not in breach of copyright as the

EIDC will be publishing the paper openly to the public. This requires a legal understanding of the journal’s T&Cs.

In these cases where the content of the research paper adequately describes the dataset the most efficient process is to copy the content that describes the dataset into a stand-alone document to be supplied to the EIDC. This extract should not constitute a copy of the journal paper.