



Draft Data Peer Review Guidelines for GDJ

Jonathan Tedds, Becca Wilson (Leicester), Sarah Callaghan (BADC),
Fiona Murphy (Wiley), Rebecca Lawrence (F1000R), Geraldine
Stoneham (MRC), Elizabeth Newbold (BL), Rachel Kotarski (BL),
Matthew Mayernik (NCAR), John Kunze (CDL), Angus Whyte (DCC),
Simon Hodson (Jisc) and the #PREPARDE project team

jat26@le.ac.uk @jtedds

Senior Research Fellow, D2K Data to Knowledge (University of Leicester)

PI PREPARDE <http://www.le.ac.uk/projects/preparde>



Data Review Guidelines for GDJ

- Panel of data scientists to be called on by the journal to conduct data reviews.

Recommend that these are scientists in the broader field of research of submitted data, they would then have the expertise to assess the quality of the data.

- Data must be submitted to an approved repository.

Data held within an approved repository implies that it meets certain data archive standards e.g. data format, null values flagged. Whilst this is true for national archive centres is this true of repositories like figshare etc.

Data Review Guidelines for GDJ

- If data is not in a national data archive/data centre the data reviewer must perform additional data checks which may include data validation.

Check the data is in an appropriate format and follows the standard as per field e.g. units, standard methodology for getting data etc. Perhaps some automated tools could be built to validate data and flag suspect data?

- Summary of the data is required from the author.
e.g. temporal coverage, spatial coverage, type of data etc should be provided so readers of the article/dataset can find out what data is included. This will be useful for people looking for datasets.

Data Review Guidelines for GDJ

- Authors required to submit specific statistics regarding the data in order for the reviewer to assess the quality of the data. These statistics will be subject specific.
e.g data ranges (are they within the norm?), standard deviation etc. Statistics required could be determined by the assigned data reviewer.
- Provide data reviewers with a checklist of criteria to include in their review.

A Checklist for Data Reviewers

- Does the dataset have a permanent identifier?
- Does it have a landing page (or README file or similar) with additional information/metadata, which allows you to determine that this is indeed the dataset you're looking for?
- Is it in an accredited/trusted repository?
- Is the dataset accessible? If not, are the terms and conditions for access clearly defined?
- Are the access terms and conditions appropriate?
- Is the format of the data acceptable? Is it in a standard or modern data format?
- Does the format conform to community standards?

A Checklist for Data Reviewers

- Can I open the files and view the data? (If not, reject straight away)
- Is there information about any proprietary software required to open the data including version number?
- Is the metadata appropriate? Does it accurately describe the data?
- Are there unexplained/non-standard acronyms in the dataset title/metadata?
- Is the data calibrated? If so, is the calibration supplied?
- Is the data flagged? Is there an appropriate description of the flagging parameters?

A Checklist for Data Reviewers

- Is information/metadata given about how/why the dataset was collected? (This may be found in publications associated with the dataset)
- Are the variable names clear and unambiguous, and defined (with their units)?
- Is there enough information provided so that data can be reused by another researcher?
- Is the data of value to the scientific community?
- Does the data have obvious mistakes?
- Does the data stay within expected ranges?
- If the dataset contains multiple data variables, is it clear how they relate to each other?