

Contributing guidelines

Thank you for considering to contribute data to the material intensity database project.

Data contributors must ensure that their inputs do not infringe any intellectual property or copyright agreements.

Data & variables

Please also refer to the and the data descriptor article in Nature Scientific Data. The database contains *primary* and *secondary* attributes. Primary attributes are mandatory throughout database and for any data new data contributions, while secondary ones are optional.

In order to make a (data) contribution to this repository please consider the following guidelines:

1. Primary data must be provided (i.e. `required == TRUE` as defined in the [codebook](#)) and cannot be left empty.
2. Provision of secondary data columns is optional, but encouraged.
3. Adding the data source to the [BibTeX reference file](#) is also optional.
4. Adhere to the given CSV [data format](#)
5. Create a [pull request](#). It is also possible to [edit files online on github](#), which may be inconvenient for larger changes.

Data can be converted or aggregated. Any alteration to the original data must be documented in one of the designated comment columns. We encourage the use of the [liberated_data project](#) if you are extracting data from non-portable, such as figures or tables in PDF files.

Data format

The file must be compliant to [RFC 4180](#).

We require that pull requests adhere to the following guidelines (most based on RFC 4180):

- The delimiter is a comma.
- One record / data point per line.
- Every record has the same sequence of fields.
- Include a header column.
- Text should always be wrapped by double quotes. Any field may be quoted (with double quotes). Otherwise fields containing a line-break, double-quote, or commas will cause data corruption.
- A (double) quote character in a field must be represented by two (double) quote characters.
- Use MS-DOS-style lines that end with (CR/LF) characters.
- Use UTF-8 encoded files.

Please refer to the [Wikipedia article](#) on CSV for more details.

Database structure

If you intend to make or propose changes to the database structure, such as adding new variables, please keep in mind to update the [codebook](#), precisely explaining the function and format of new columns. Also, please keep in mind that the codebook is intended to be validated by a Python script (*UPDATE required*).

References

[references/](#) contains the references to the data in [data/](#). Adding a reference for a new data point is optional. Please use the [BibTeX format](#) and append the .bib file at the end with the new reference. Please use only the common fields, such as author, year, etc. BibTeX references can easily be exported from Google Scholar (by clicking on the parenthesis icon on the article), Mendeley (Menu Edit > Copy As...), or zotero.

Tools

Since many users will probably use Microsoft Excel to process the data, the requirements for formatting account for some of the Excel quirks. For instance, the Boolean variable `True` will always be converted to uppercase by Excel. This could be avoided by a leading single quote `'`, but Excel keeps replacing the value on import. Therefore, the uppercase values `TRUE` and `FALSE` are being used. Other than that [Microsoft Excel seems to do a decent job of exporting RFC compliant csv data](#).

Please consider also contributing to the [liberated data](#) project in case you extract data from non-portable formats such as figures in PDF files. The project facilitates transparency and reproducibility in science.