

Dataset documentation outline

DRAFT, Leigh Dodds, 4th February 2016

The open data maturity model recognises the importance of publishing well-structured, high quality documentation and metadata to accompany data releases.

This document has some DRAFT guidance that provides some further detail for data publishers intending to improve their documentation process.

Metadata

- The ODI recommends use of standard metadata vocabularies such as [DCAT](#) to help describe datasets and the distributions. The [Open Data Rights Statement vocabulary](#) also identifies the types of licensing metadata that are important to share with users.
- The [Open Data Certificates](#) application provides a detailed list of the metadata fields that should accompany data releases. [High level ratings](#) have more structured metadata

Documentation

The following outline provides a basic template for dataset documentation. This is just an outline and is not intended to be exhaustive.

- **Summary**
 - A general overview of the dataset that broadly describes its contents, how it was collected, etc.
 - This will help orientate someone about what the dataset is about and help them understand whether they might be able to use it
- **Metadata**
 - This section would list the key facts & figures about the dataset, e.g:
 - when it was published and last updated
 - when it will be next updated
 - frequency of updates
 - notes on its geographic coverage
 - the time period being reported on, etc.
 - the licence for re-using the data
 - any requirements around citing/attributing the dataset
 - The metadata should be presented for human readers and may also be machine-readable
- **Structure of the data**
 - A detailed breakdown of each of the columns in the spreadsheet or CSV file. If there are multiple worksheets then these would be itemised too.
 - For each column I'd expect to see:
 - the name of the field

- a definition of what is being reported
 - what type of the field contains, e.g. a date, number, or value from a controlled vocabulary
 - any additional notes that will guide interpretation of the contents of the column.
- This section might link to additional supporting documentation about controlled vocabularies used in the data, etc.
- **Background/Methodology**
 - This provides a further level of detail for users interested in learning more about how the dataset was compiled, what external sources were used to compile it.
 - Any supporting notes that describe relevant calculations, quality control procedures, etc.
 - Suggested things to include are:
 - how and when the data has been collected
 - notes on any relevant quality control or assurance processes
 - an outline of the process by which the data was analysed in order to derive the final results
 - stating any assumptions that might influence how the data is interpreted
 - stating any known quality issues or limitations in the data, to ensure that it is correctly used to derive insights
 - notes on how the data might be used by others, to help steer reusers together useful and desirable outcomes
- **Support Information**
 - Finally, contact details identifying where someone can go to ask further questions, discuss the data, report errors, etc.