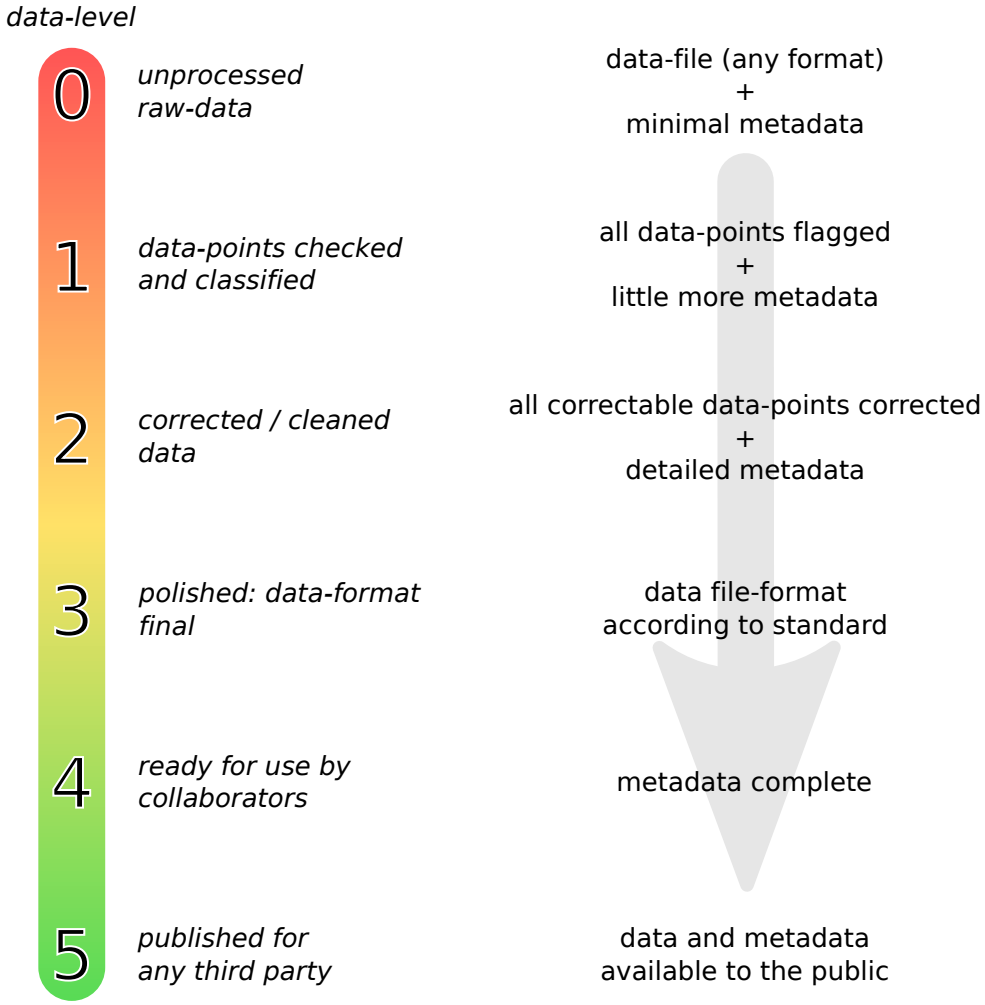
Data-Levels

Background & Motivation

Scientific data commonly undergo a series of preparative processing-steps (e.g. normalisation, outlier-detection, de-trending) ahead of their actual analyses. When sharing data, it is essential to communicate outstanding processing-steps to avoid misinterpretation of the data. This can be done via *Data-Levels*, which roughly encode the overall state of a data-set.

The general processing-state of a data-set is classified according to 6 increasing *data-levels*: level 0 to level 5. Lower levels indicate preliminary stages at which the data can not be used for scientific analyses. By sanity checking, correction, and proper formatting, the data reach higher levels, given that accompanying metadata are complete. The process is sketched in the following figure:



In general, a data-set's data-level never decreases over time, but only remains unchanged or reaches higher levels. Data-levels need not be stepped up in increments of 1, but intermediate data-levels can be skipped, if appropriate.

C. Feenders, 20th November 2018

Together, the data-level is determined by 5 criteria:

metadata Each data-set is accompanied by meta-information about how to interpret the actual data (e.g. parameter, unit, offset). It also comprises the method used to collect, check, and correct the data. While minimal information about the data (e.g. parameter measured and unit) may be sufficient for preliminary analyses, in the final stage, the metadata should contain all information needed to draft the method-section in publications.

quality-flags and **values** Individual data-points may be of very different quality for different reasons (e.g. sensor malfunction, failed calibration, high measurement error). Data points can be classified according to their quality by assigning quality-flags (see document on *Data Quality Flags* for details). Already for basic data-levels all data points must be checked and flagged correspondingly. Wherever possible and sensible, false values must be corrected to reach the next data-level.

format Data can be stored in a multitude of different file-formats (e.g. `.txt`, `.csv`, `.xls`) and even for one file-format (e.g. Excel-Spreadsheet) many possibilities to order the data within the file (data-format) exist. To ease the usability of data, it is often advisable to agree on a common file- and data-format. While the data may be in any format for lower data-levels, it must correspond to the standard (if any) at higher levels.

public Data thoroughly analysed by its originator and collaborators may still provide new insights, if re-analysed differently or if combined with other data-sets. Data should thus be placed at the disposal of the research community (e.g. supplementary information, citable database upload, specific web-site). If this and all other criteria are fulfilled, the data are assigned the highest data-level.

The following table summarises the minimum requirements for each data-level:

data-level	**definition**	criteria				
		metadata	**quality-flags**	**values**	**format**	**public**
0	raw data	minimal	none	no check	any	no
1	checked	**little**	**flagged**	**checked**	any	no
2	corrected	**detailed**	flagged	**corrected**	any	no
3	polished	detailed	flagged	corrected	**standard**	no
4	ready	**complete**	flagged	corrected	standard	no
5	published	complete	flagged	corrected	standard	**yes**

The last row for which a data-set fulfils all criteria corresponds to its data-level. It is advisable to document the data-level in the metadata and to encode it prominently in file-names.

Technical Information The data-levels in this document relate to those used for the coastal observation system COSYNA (Breitbach et al., Ocean Sci., 12, 909–923, 2016). Missing descriptions have been meaningfully filled. An additional data-level has been added to differentiate between levels *ready* and *published*.