Data.gov.sg Data Quality Guide

Data.gov.sg Data Quality Guide

Key Principles

1. Machine Readability

Machine-readable data is data in a format that can be understood by a computer. This means that relevant fields can be extracted and parsed by without human input.

Just because a dataset is *digitally accessible* does not necessarily mean it is *machine readable* as well (more on that here).

In the spirit of open data, datasets must be provided in non-proprietary file formats. Tabular data is provided as CSV files or in JSON format.

Metadata must be excluded from data files. They will be included in a zip file with the data file (see the Data Package section below).

2. Tidy Data

The concept of *tidy data* is based on this paper by Hadley Wickham, author of popular R packages ggplot2 and plyr.

Hadley proposes three principles:

1. **Each column is one variable.** This means each column must have the same unit of measurement

2. **Each row is one observation.** Note that this does not mean there is only one observed variable per row. There can be multiple observed variables per row: “An observation contains all values measured on the same unit (like a person, or a day, or a race) across attributes.” (Wickham Pg 3)

3. **One type of observation unit per table** For example, observations on fertility rates of the whole female population and that females of different age groups should be in two tables:

    - “Total Fertility Rate”, where observation unit is the entire female population
    - “Total Fertility Rate by Age Group”, where observation unit is female population of different age groups

3. Consistency

Each variable should be processed and documented in the same way across datasets. Examples include:

- Date formats
- Null and negligible values
- Unit of measurement

This makes it easier for users to mashup and use multiple datasets, even from different agencies

4. Granularity & Precision

As far as possible, agencies should provide raw, granular data instead of aggregated and process data, such as percentages.

Totals and sub-totals should be in separate tables if needed. For example, there are cases where aggregate numbers (e.g. totals, indices) cannot be derived from granular data points.

5. Human Readability

Lastly, datasets should also be presented in a way that makes sense to human users.

This is why we have created a new set of metadata specifications along with this Data Quality Guide, to ensure that datasets are documented clearly. For instance, each table must be accompanied by a schema, which states the variable type.

Plain language should be used in naming datasets and the elements in them as far as possible. Column names should be descriptive and meaningful. All jargon must be explained clearly, and where necessary, reference links should be provided.

Data Packages

We adapted this concept from Open Knowledge International's Data Package specification, which allows us to package metadata with data files.

We chose to provide metadata in YAML instead of JSON format, as YAML is both human- and machine-readable.

When users download a dataset from **Data.gov.sg**, the zip file contains:

- A metadata text file (in YAML syntax)
- The resource file(s)

Structure for Tabular Data

1. Column Headers

- Only alphanumeric or these 3 special characters: .-_
    - Ampersand (&) should be replaced by "and" if needed
- Each must be unique
    - Can't have two headers called "duration"
- All headers are in lower case
    - e.g. "rating"
- If header contains more than one word, use underscores to join

        - e.g. vehicle_type
- Units of measure should be omitted
- Keep short (less than 25 characters)
        - The full name can be stored in schema section of metadata file

2. Column Order

- *Date and time variables* should always be in the first column for time series data
- *Fixed variables* should be ordered with the highest-level variable on the left and most granular variable on the right
- *Observed variables* should always be on the rightmost columns

3. Row Order

- Rows should be ordered from the leftmost column to the rightmost one
- For datetime variables, order chronologically
- For fixed variables, order alphabetically

4. Date and Time Variables

- Based on ISO8601, an international standard for representing date and time. We chose the “extended format” with the hyphens because it is humanly more readable.
        - Compare 2016-01-01 to 20160101
- All date and time variables must be in UTC +8hrs unless specified.
- Date variables:

Interval	Column name	Format	Range of values	Example
Annual	year	YYYY	YYYY: 1900 onwards	2015
Monthly	month	YYYY-MM	MM: 01 to 12	2015-01
Daily	date	YYYY-MM-DD	DD: 01 to 31	2015-01-01
Weekly	week	YYYY-[W]WW	[W]WW: W01 to W52	2015-W01
Quarterly	quarter	YYYY-[Q]Q	[Q]Q: Q1 to Q4	2015-Q1
Half-yearly	half_year	YYYY-[H]H	[H]H: H1 or H2	2015-H1

- For financial periods, prefix “financial_” to column name:

Interval	Column name	Format	Example
Financial, annual	financial_year	YYYY	2015

Financial, quarterly	financial_quarter	YYYY-[Q]Q	2015-Q1
Financial, half-yearly	financial_half_year	YYYY-[H]H	2015-H1

    - Financial year start-date must be indicated in metadata

- For date-time variables:

Type	**Column name**	**Format**	**Example**
Date + time	date_time	YYYY-MM-DD[T]hh:mm	2015-01-01T12:00
		or YYYY-MM-DD[T]hh:mm:ss	2015-01-01T12:00:00
Time only	time	hh:mm	12:00
		or hh:mm:ss	12:00:00

- Specify the timezone if it is not UTC +8hrs:

Type	**Column name**	**Format**	**Example**
Date + time	date_time	YYYY-MM-DD[T]hh:mm+hh:mm	2015-01-01T12:00+00:00
		or YYYY-MM-DD[T]hh:mm:ss+hh:mm:ss	2015-01-01T12:00:00+00:00:00

5. Textual Variables

- UTF-8 encoding should be used
    - This ensures that special characters such as Chinese characters can be decoded by users
- No line breaks within cells

6. Numeric Variables

- No commas
    - E.g. “1000” instead of “1,000”
- No units of measurement
    - Units should be in metadata instead
- Express as full number where possible
    - If rounded, indicate in metadata
    - E.g. “1200000” instead of “1.2” (million)
- No rounding if possible
    - Give raw numbers as far as possible
    - If rounding is needed, try to provide at least 2 decimal places of precision

- Percentages can be expressed as either a proportion out of 1 or 100.
    - E.g. 20% can be expressed as 20 or 0.2
    - The representation of percentages must be consistent throughout each CSV file
    - Agencies must indicate how percentages are expressed in the schema

7. Location variables

- Coordinates in EPSG 4326 or EPSG 3414:
- Should be represented in two columns
    - EPSG 4326: latitude and longitude or
    - EPSG 3414: x_coord and y_coord
- In positive/negative floating point
    - e.g. latitude: 1.2896700; longitude: 103.8500700
- EPSG should be indicated in metadata
- Addresses should be represented in two columns if possible
    - address: e.g. 1 Sims Avenue, Singapore 123456
    - postal_code: e.g. 123456

8. Null/negligible values

- For any variable type:

Value	Meaning
na	Datum not available or not applicable
-	Datum is negligible or not significant
s	Datum is suppressed

- If possible, explain why there are such values in the metadata

Reserved column names

The following column names should be used only if they adhere to the definitions in this guide:

- year
- half_year
- quarter
- month

- week
- date
- datetime
- time
- financial_year
- financial_half_year
- financial_quarter
- latitude
- longitude
- x_coord
- y_coord
- postal_code

Contact us

We will continue to review this Data Quality Guide to ensure that we meet the needs of our users.

We welcome all feedback to improve the quality of data published on **Data.gov.sg**. Raise an issue on Github or drop us an email at feedback@data.gov.sg if you have any comments or queries on the guide.