# GENERAL GUIDELINES TO GUARANTEE THE QUALITY OF OPEN DATA

### EVADE NON-PROCESSABLE DATA FORMATS

Share data with **reusable open-format** files that facilitate data access.

### USE A STANDARDIZED CHARACTER ENCODING

It is recommended to use an internationally recognized, standardized or used **character encoding**, such as **UTF-8** encoding.

### NAME THE COLUMNS APPROPRIATELY

Use only lowercase characters. The fields and specifications must be collected in the data dictionary that documents the dataset. No special characters, titles or punctuation marks should be used either. Spaces must be replaced by dashes.

### AVOID MISSING VALUES

To avoid confusion, **absent values should be clearly marked as null values (NA).**

### SHUN DUPLICATE RECORDS

Standardize the collection of data and its storage, centralizing the process in a single information system, so that duplicates are easily detectable and can be automatically eliminated.

### STANDARDIZE DATA VALUES

To standardize the structure and values of the fields, it is recommended to use **reference vocabularies**. The structure must be documented in the **data dictionary**.

## DATA QUALITY ATTRIBUTES

**Consistency**

**Accessibility**

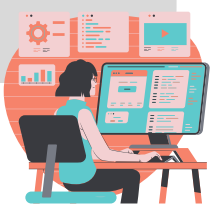
**Precision**

**Comprehensibility**

**Traceability**

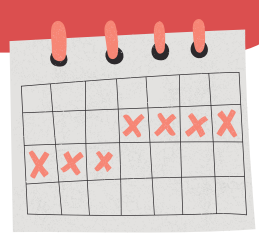
### PROVIDE AN ADEQUATE AMOUNT OF DATA TO FACILITATE ANALYSIS

Ensure that a **reasonable amount of data** is published so that there is enough context and that users can derive value from its exploitation.



### FORMATTING DATE AND TIME VARIABLES

Dates should always be encoded using the ISO standard, yyyy-mm-dd for the date and hh:mm:ss for the time.



### FORMATTING NUMERIC DATA

Use a point as decimal separator (internationalization). Avoid thousands separators. Negative values with sign (-). In columns with integer values, do not use decimal separators or mix text with numeric values.

### AVOID MIXING NUMERICAL SCALES

**Try not to change the scale** over time. If necessary, provide the data at both scales and document the change of scale.

### ELUDE MIXING RANKS IN THE SAME DATA SET

Publish the data with the **highest level of disaggregation**. If not possible, maintain consistency across all variable values.

### INCORPORATE VARIABLES WITH GEOGRAPHIC INFORMATION

Post the data with **geographic coordinates in two independent columns**: "latitude" and "longitude".

### AVOID THE INCLUSION OF SUBTOTALS, TOTALS OR GROUPINGS

Present the **highest possible level of disaggregation** of the data it contains.

### SKIP DATA FRAGMENTATION AND DIFFICULT LOCALIZATION

Improve the **organization and labeling** of the contents, and you need to establish connections between the different sets of data.

*Source: "Guía práctica para la mejora de la calidad de datos abiertos: Secretaría de Estado de Digitalización e Inteligencia Artificial del Ministerio de Asuntos Económicos y Transformación Digital" (September, 2022).*
*https://datos.gob.es/ca/documentacion/guia-practica-para-la-mejora-de-la-calidad-de-datos-abiertos*

ICAC Institut Català d'Arqueologia Clàssica

Creation date: 2023-10-17

openscience.icac.cat