Arcus Research Data Management
Best Practices for Data Dictionaries
2020 May 15
arcus-support@email.chop.edu

Best Practices for Data Dictionaries

A data dictionary will help people better understand the scope, purpose, and nuance of the data you are collecting. Some data dictionaries are extensively detailed, but even a basic minimal data dictionary is better than none at all. Data dictionaries are usually used for tabular datasets, but can be used for data in other formats as well! Below are some reasons to consider making a data dictionary, recommendations about where to start, and options for extending your documentation past the basics if you so choose.

Purpose

Why create a data dictionary? Many think a data dictionary is only necessary if sharing data with external parties. However, there are many good reasons to create one even for internal use. These include (but are not limited to!):

- **Internal references for consistency**: our ideas of why and how we collected or created data can shift over time. A data dictionary provides a reference point to maintain consistency during the course of long research and analysis efforts
- **Onboarding**: a data dictionary is a good way to help new people come up to speed
- **Protect against knowledge loss**: creating documentation such as a data dictionary ensures crucial knowledge is maintained even during staff turnover
- **Troubleshooting**: if there are questions or issues with the data, a data dictionary can be a useful first point of reference to start troubleshooting

Format

To the extent possible, a data dictionary should be maintained in a structured format. The most common structure is a csv file, but other options include YAML, JSON, or any other structured data format you are comfortable with.

Content

Below are the fields you can consider including in your data dictionary. Only a few are considered truly “required” - the rest are optional but can be extremely helpful, so you should consider whether they make sense to collect in your case. You may also have additional fields to include that are not listed here; you know your data best!

Name: (required) provide the name of the data element you are describing as it appears in the dataset.

Description: (required) provide a brief description of the data element. Things to include here if applicable: source of data element, units of measure, formulas for calculated fields, nuances of data capture environment, anything else relevant to understanding what this field means and how to use it

Arcus Research Data Management
Best Practices for Data Dictionaries
2020 May 15
arcus-support@email.chop.edu

Human-readable Name: (optional) provide a human readable name / title for the data element. This can be handy if the names in your dataset are hard to parse or hard to understand.

Type: (optional) indicates the type of data i.e. numeric, string, date, etc. If you are collecting data in a database that enforces types, sometimes this can be easily extracted. This is useful to know for future transformation or integration of datasets

IsNull: (optional) this a yes/no field that indicates whether the item can be null (absent of information) or not. If this is set to “no”, this indicates data should *always* be present in the field, and is helpful to users who may wonder whether an absence of data is to be expected or indicative of a problem

Values: (optional) if there is a restricted list of values that can populate this field, include that here. An example might be “eye color” with the value list “blue, brown, grey”. Providing the list of values helps users understand why data may be absent. E.g. are there no green-eyed people in the dataset because there were none in the population or because “green” was not one of the available values for this field?

Date added: (optional) indicate the date on which this particular field was added to the dataset. This can be useful for understanding why data may be absent.

Date removed / Date deprecated: (optional) indicate the date on which this particular field ceased to be collected as part of the dataset. This can be useful for understanding why data may be absent.

Relationships: (optional) indicate whether the data element is related to other data elements in your data set (in relational database terms, this is where you would indicate a “foreign key”)

Additional Documentation

In addition to documenting data fields, your data dictionary should also include an overarching description of the dataset itself. If you have more than one dataset, or more than one data table, include a description for each. The descriptions for data tables and datasets should include some idea of the scope of data included, as well as the purpose in collecting or creating the data and its intended use.