Data Dictionary Format

Last Updated: 15 September 2023

|  |  |  |  |
| --- | --- | --- | --- |
| COLUMN | DESCRIPTION | POSSIBLE VALUES | REMARKS |
| NAME | Name of the column | gender, dob |  |
| DESCRIPTION | Short description of the variable. |  |  |
| TYPE | General data type. Specifically, how it appears superficially. | numeric, string, date | Data can often appear in multiple formats. For e.g., categorical data can often be saved as ordinal integers, Boolean, which appear numerical, or as text, e.g. ‘YES’ or ‘NO’, or simply as unique sequence of numbers, e.g. ‘3828’ and ‘4271’ (which looks numeric but are actually strings). Another example is when dates are saved as strings, or as date-formats using excel, or its numerical equivalent. These two columns help data users navigate this confusion and facilitate development of automated scripts.  TYPE indicates how the data appears superficially, whether it should be processed as a number, string, or date when loading or saving the data.  SPECIFIC DATA TYPE indicates how the data should be processed in the ideal situation during analyses, for e.g. a datatype such as education level should ideally be ordinal, though it can be treated as categorical, ordinal, or int. |
| SPECIFIC DATA TYPE | Specific data type. Specifically, the ideal way of processing the variable. | E.g. 1: If numeric, specify as ‘float’, ‘int’, ‘boolean’, ‘ordinal’, ‘categorical’, ‘date’ etc.  E.g. 2: If string, specify as ‘categorical’, ‘ordinal’, ‘free text’ etc.  E.g. 3: If date, specify as ‘date’ |
| CODINGS |  | E.g. 1: If TYPE is ‘date’, use excel convention to indicate date format, e.g. dd/mm/yyyy, mm-dd-yyyy, etc.  E.g. 2: If SPECIFIC DATA TYPE is ‘boolean’, ‘ordinal’, ‘categorical’, specify exhaustively all possible entries, delimited by ‘; ’, e.g. YES; NO; N.A. OR Male; Female, OR 1; 2; 3; 4; 5  E.g. 3: If TYPE is ‘numeric’, specify range. E.g. [0,100] OR (3,4). | Take special notice of capital/small letters to avoid confusion.  TIMS might have specific date conventions, follow as convenient. |
| FREQUENCY | For longitudinal data. Use to indicate if the variable is collected during a particular visit type. | E.g. 1: BASELINE; 6 WEEK; 6 MONTH  E.g. 2: VISIT 1; VISIT 2 | Leave blank if not longitudinal data. |
| CATEGORY | Use to group the variable under a specific category. | E.g. 1: DEMOGRAPHICS  E.g. 2: ECHO  E.g. 3: LIFESTYLE VARIABLE |  |
| SECONDARY | Whether the variable can be derived from other variables present in the dataset. | E.g. 1: If yes, ‘Y’  E.g. 2: If no, ‘N’, or leave blank. | For instance, BMI is a secondary variable if it is computed from ‘height’ and ‘weight’, and the two variables are also included in the dataset. Other examples are ‘age\_decade’, where subjects of ages between 30 to 40 are grouped together as ’30-40’ to reduce granularity of the variable, or ‘dementia’, a diagnosis derived from answers to questions also present in the dataset.  If yes, explain how the variable was computed from other variables, bmi formula, diagnosis standard/criteria etc, either in the CONSTRAINTS or REMARKS column. |
| CONSTRAINTS | How the variable is dependent on other variables. | E.g. 1: ‘head\_circ’ (head circumference) is a variable collected for ‘age’ <= 6. Leave empty if ‘age’ > 6.  E.g. 2: collected only for data cohort ‘<COHORT NAME>’ or hospital ‘<HOSPITAL A>’.  E.g. 3: ‘ever\_pregnant’ only collected for females, above age of 12. If ‘male’ or ‘female’ below age of 12, recorded as ‘N.A.’ If ‘female’ above age of 12, either ‘YES’, ‘NO’, or ‘UNK’.  E.g. 4: ‘BMI’ only computable if ‘height’ and ‘weight’ are also collected. Leave blank if either values are blank. | This information will help data users decide if a value is missing/unknown (should be collected but not collected), or not applicable (not collected because of procedure).  Note that the value of a variable might be dependent (or conditional) on other variables, but it’s not necessarily derived from other variables; CONSTRAINTS and SECONDARY are complementary, but the former does not imply the latter. |
| REMARKS | Additional comments, such as how the data is encoded, and/or concerns related to the variable. | E.g. 1: How categorical variables are encoded as integers: 1=NO, 0=YES, -1=N.A.  E.g. 2: sensitive OR self-reported variable, etc.  E.g. 3: metric unit used for collection, ‘cm’, ‘m’, ‘inches’, etc. | It is often necessary to leave a note to remind data owners/users of the difficulties encountered during data collection, the corresponding response, and associated concerns. Some of these remarks can be included in the variable description, or here, if they are deemed miscellaneous. |