Guideline for Describing a Data Dictionary

Authors: Martin Bialke, Peter Penndorf, Daniel Fredrich, Kerstin Weitmann

English translation by Henriette Rau

English Version 1.3, 04.08.2017

mosaic-project@uni-greifswald.de

Content

NOTICE

You are using this template for your own research? Please cite the MOSAIC project in your work and manuscripts:

Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T , et al.*
MOSAIC. A modular approach to data management in epidemiological studies.
METHODS OF INFORMATION IN MEDICINE. 2015; 54(4):364-371.
http://dx.doi.org/10.3414/ME14-01-0133

1 Motivation

In order to answer the research question(s) of a study or registry, information is collected - using variables. Variables should always be atomic, and represent, inter alia, responses of a questionnaire, measured values or influencing factors (e.g. examiners, devices, measuring methods or time stamps).

Before starting a registry or study, the study plan must specify clearly which variables are required to answer the research question (see Good Epidemiological Practice (GEP), Guideline 3.5 [1]). Additionally, the comprehensive description of these variables should be carried out with a focus on the subsequent data collection, quality assurance, and evaluation. Typically, the result of such a formal description is a Data Dictionary. However, this formalization can pose a challenge to epidemiological researchers without excellent technical knowledge.

Why is a Data Dictionary needed?

A Data Dictionary supports researchers to document data, which should be collected, in a comprehensive and consistent way. It is not possible to analyze or exchange data (e. g. with other researchers) without additional information such as definitions, characteristics, validity ranges, context of data collection, units of measurements or coding of variables.

The aim of this guide is to help epidemiologists and scientists to develop a Data Dictionary as briefly and precisely as possible. For this reason, aspects, which should be considered, are presented and recommendations for the procedure are given.

Why is completeness and correctness of the Data Dictionary from the very beginning a key to success?

The definition of the Data Dictionary must be carefully prepared and coordinated, as it is the basis and starting point for all subsequent steps in the course of the study or registry. If changes to a study's data set or the Data Dictionary are required after the start of the study or registry, they have considerable organizational and timeline-related effects - the effort of such changes is oftentimes underestimated.

After detailed description and documentation of a change, the data set of a study or registry has to be adapted and reconciled in terms of structure, dependencies and characteristics with all the partners involved. If the process of data collection changes, survey forms have to be revised and databases have to be updated. This requires a re-examination of the system with further software tests. Additionally, existing standard operating procedures (SOP) must be adapted, which can result in retraining and re-certification of study staff. Moreover, if a variable is changed, it must always be carefully checked to which extent the survey context is maintained. For example, a changed measuring method of a variable affects the comparability of existing and future values. Consequently, this makes data analysis more difficult.

2 Prerequisites

Before a Data Dictionary can be developed, the research question of the study or registry has to be defined. Also, required variables and parameters should already be known in principle, e. g. units, dependencies, possible confounders or disturbance variables [2]. If available, it is recommended to use already existing national and international data sets [2].

"*All variables of interest should be precisely defined and operationalized as much as possible according to professional standards. Measurement and survey instruments that are as valid and reliable as possible are to be used*." (Excerpt from the Good Epidemiological Practice, guideline 3.5, p. 12, *[1]*)

3 What should be included in a Data Dictionary?

3.1 Defining variable names

Literature and practical experiences give different recommendations as to how variable names should be chosen. Simple identifiers including the question number with ascending numbering are opposed to self-explanatory variable names that relate to the questionnaire and content of the variable (see [3], p. 27ff). In general, the following recommendations can be summarized for the variable name (short name):

***Use descriptive identifiers for variable names**. The content of the variable should be comprehensible without additional documentation and knowledge of the context.*

***Use short and long names**. The short name must be unique and is used as column name within the database. The long name describes the variable more precisely and is usually used in forms as a description field or label. In addition, help texts as well as detailed variable descriptions can be presented.*

***Use prefixes**. Prefixes generally support the clarity and unambiguous identification of a variable within the Data Dictionary. For example, a variable with form prefix can be directly assigned to the source form, even if multiple forms are used.*

***Use suffixes**. Suffixes should be used when a plurality of descriptive identifiers are needed for similar variables as part of an enumeration (e.g., risk factor_1, risk factor_2, etc.). Also, suffixes can be used as references to data types or units.*

From a practical point of view, the later data processing must be taken into account when defining variable identifiers:

***The length of the variable name must be based on current standards.** Please keep in mind possible limitations of the database system [4] and the analysis software.*[1]

***Avoid upper case.** To avoid problems with the processing of the collected data across operating system boundaries, use only lower case.*

[1] For example, variable names in SPSS are allowed to have a maximum length of 64 characters only since version 12.0. Previous versions were limited to 8 characters. [5]

***Avoid reserved words**. Databases and many other technical systems use reserved words such as "begin", "date", "table", "system", "name", etc. If these reserved words are used as variable names, unwanted side effects during data processing as well as system errors may occur.* [2]

***Avoid umlauts**. Umlauts - above all in the German language (e. g. ä, ü, ö, etc.) - cause most often problems with the storage and analysis of data.*

***Avoid blanks and special characters**. Variable names, which contain spaces, dots and or special characters (e. g. "-", "%", "$", "&", "/", "(", ")"), already complicate creating a database, since these characters may be invalid depending on the database system [4], and cause system errors. However, connecting words within a variable name using underscores ("_") is possible.*

***Enable international dissemination**. If possible, use English variable names to support understanding and usefulness of the variables for international analyses and publications.*

***Provide structure information**. Names of questionnaire variables should contain references to the questionnaire structure (chapter or section numbers). This structural information is required when only parts of the collected data sets are exported for data analysis.*

Table 1 Examples for "bad" variable names

Variable name	*Problem*	*Improved variable name (short name)*
Waiting-time (h)	Upper case, special characters, blanks	frm_admission_waitingtime_hours
n1, n2, n3	No reference to variable content	frm_admission_accidentcause1, frm_admission _accidentcause2, frm_admission _accidentcause3
date1	Not descriptive enough	frm_ admission _date

3.2 Defining variable characteristics

3.2.1 Determining data types

If required, nearly every to be collected variable can be represented by basic data types:

***Integers** are particularly suitable for the specification of frequencies as well as the representation of categorical data, e.g. coded selection possibilities for the cause of an accident.*

***Floating point numbers** are particularly suitable for metric values, e.g. the measurement of blood alcohol concentration. The required decimal places should always be specified.*

***Character strings** are suitable for texts. The maximum length of a string should always be specified, e.g. "description of the accident (max. 255 characters)".*

***Date and time** can be displayed differently depending on the data capture tool and database system. Basically, the usage of character strings is possible. In any case, the format to be used for date and time should be uniformly defined, e.g. "hh:mm" or "dd.mm.yyyy".*

[2] For a list of reserved words using the example of MySQL see the following: http://dev.mysql.com/doc/refman/5.6/en/keywords.html

3.2.2 Defining ranges of values

For each variable, a valid range of values should be defined to minimize the risk of entering implausible values during data collection. The values range must include all realistic values and must not be too narrow.

If possible, specify the scale of measurement[3] for future quality assurance and analyses. This allows the distinction of values ranges and, thus, supports the data analysis. For example, the variable "accident cause" (values range 1..5 integer, nominal) is analyzed in a different way than a variable "body size", which can also be specified as an integer, but is metric and, thus, allows a comparison of values and mean values.

Table 2 Examples of variables and possible values ranges

Objective/ Variable	***Possible values range***	***Scale of measurement***
Selectable options 1,2 and 3	1,2,3	nominal
measurement in percentage	0 - 100	ordinal
Height in cm	1 - 299	cardinal
Admission date	01.01.2015 - 31.12.2015	cardinal

3.2.3 Coding valid values

If questions are answered by selecting one or more predefined answers or response categories, each possible answer has to be encoded unambiguously. The coding must be clear, extensible and uniform, represent and mutually exclude all possible answers. For this purpose, numerical and alphanumeric codes are used [3].

Table 3 Examples for coding valid values. In these examples the assignment of value and coding is based on the order of values and codings.

Objective	***Valid values***	***Possible coding***
Standard question	No, Yes	0,1
Gender	Male, Female	m, f
Accident cause	scalding, flame, explosion, acid, alkaline solution	1,2,3,4,5

Further examples and tips can be found e. g. in the *Guidelines for research data management* [German, original title: Leitlinien zum Forschungsdatenmanagement] [3] (p. 29ff).

3.2.4 Specifying qualitative missings

In general, missing information (missings) should always be described as precisely as possible. If this is neglected, it is not possible to make a statement in the subsequent analysis of the collected data records as to whether a variable was incorrectly recorded or the missing is caused by e. g.:

- missing information (respondent does not know the answer)
- lack of willingness (respondent refuses to respond)

[3] More about scales of measurements: http://www.reiter1.com/Glossar/Skalenniveaus.htm [German], Accessed: 15.09.2015)

- not able to gather information due to errors in the survey (question / answer option not applicable)
- question not asked / test was not carried out
- missing documentation (result was not recorded)

In order to be able to give qualitative statements about missing values and, thus, to prove the completeness of data for statistical analyzes and reports [2], all missings must be encoded.

Ensure uniformity and consider the ranges of values for valid values when coding the qualitative missings. The *Guidelines for research data management* [German, original title: Leitlinien zum Forschungsdatenmanagement] provide numerous tips and practical examples (see [3] p. 31ff):

***Use codes for missings outside the valid values ranges**: Use the highest numeric codes that can be displayed outside the valid values ranges (e. g. 99977 = not asked, 99978 = no answer given, etc.) or use negative codes for missings (e. g. -1, -2, etc.).*

Alternatively, the coding of qualitative missings can be based on *HL7 Null Flavour Standards* [5] or the *Coding of Null Values in OpenClinica* [6].

3.2.5 Checking for dependencies

If a variable is only allowed to be gathered if another variable has a specific value, this dependency and the resulting condition must be clearly defined.

Table 4 Examples for dependencies of variables

Variable A	***Variable B***	***Dependencies***
frm1_accident_cause_11 (accident cause: other)	frm1_accident_description (description of accident's context)	The description of the accident's context must only be given, when "other" was selected as accident cause.
frm1_krea_mgdl (creatinine value in mg/dl)	frm1_krea_umoll (creatinine value in µmol/l)	The simultaneous entry of creatinine in both units (mg dl and µmol/l) is not permitted.
frm1_admission_date (admission date)	frm1_discharge_date (discharge date from hospital)	The discharge date must be the same or a later date than the admission date.

Typical examples are gender-specific issues: Men do not need to answer questions about pregnancy and can "skip" corresponding questions in the questionnaire. These jumps are marked by jump variables in the Data Dictionary. By using jump variables, the requirements of the subsequent analysis, quality assurance, and plausibility check can be taken into account and supported during data collection.

3.2.6 Identifying calculable values

Automatic calculations of variables, such as the body mass index (BMI) as a value derived from the height and weight of a person, can reduce the burden on the survey and the inconsistencies that occur due to interpersonal calculation errors. However, incorrect formulas can affect the quality of the data systematically. Thus, the need for checking and validating the entered data remains.

***Identify calculable values:** Based on the dependencies between variables, you can specify the needed formulas.*

3.2.7 Defining mandatory variables

Basically, the data set of a study should only include all variables necessary to answer the research question of the study or registry. Such mandatory variables form the Minimal Data Set (core data elements [2]).

If further variables are gathered, mandatory and optional data should be clearly distinguished within the study's data set. However, one risks incomplete data entry by using optional variables, which may lead to problems during subsequent data analyses [2].

3.2.8 Specifying units

In order to reduce the error potential during data analysis, all units of a variable must be specified unambiguously. Therefore, the following aspects must be considered:

***Use common units**: Measured values are recorded in different units depending on the context during data entry. When specifying the unit of a variable, use units from common practice and standards, e.g. international SI units (in the current version as amended) [7].*

***Please ensure uniformity of the data**: If possible, similar variables should be represented in the same units, e.g. size and length data should always be gathered in "cm". Avoid varying information such as "%" and "percent".*

4 Best Practices

4.1 Using standards as references

According to the GEP [1], national and international standards should be used when describing variables and their characteristics if applicable. Thus, the quality and comparability of data from epidemiological studies and registries is supported.

Within epidemiological studies and registries well-established standards for classification of diagnoses and diseases (e. g. ICD [8], DIMDI), for describing variables and their characteristics (e. g. LOINC [9]) as well as for describing metadata (e. g. National Metadata Repository [10]) can be used as references.

4.2 Identifying and communcating obscurities

The definition of variables requires a significant amount of communication between all parties involved regarding the necessity and appropriateness of the variables, ranges of values, data types and dependencies - even after good preparation and planning.

To reduce the additional expenditures, which were already mentioned in chapter 1, originating from belated changes and adaptations of the Data Dictionary, emerging obscurities and problems during the development of the Data Dictionary should be documented, communicated and solved in collaboration with all involved parties.

4.3 Revising possible answer options

Avoid multiple-choice answers for a question, e. g. to name multiple accident causes. These answers are mostly saved as concatenated data elements (e. g. separated by comma) in the database and

complicate data analysis as well as reduce the quality of gathered data. Revise the question if necessary.

***In order to analyze data effectively, a variable should always contain only a single value**. If necessary, use further variables for each answer option (e. g. frm1_accidentcause1, frm1_accidentcause2, etc.) and enter whether this variable is "true" or "false".*

Additionally, avoid the entry of not-categorized free text. Free texts increase the effort for assuring and analyzing data significantly, since they cannot be analyzed automatically.

5 Approving the Data Dictionary

The description of the Data Dictionary, which is based on this guide, should be used as basis for conceptualizing forms for data capture[4] and the required data model.

In order to avoid an unnecessary amount of work, as many researchers as possible, including methodologists, should be consulted to control the quality of the developed Data Dictionary. It has to be examined whether the data set has a sufficient depth of analysis and quality [2] and whether a scientifically correct answer to the study's / registry's research question can be ensured.

The workgroups of the *German Society for Epidemiology (DGEpi),* the *German Network of Health Services Research (DNVF),* and the *Technology, Methods and Infrastructure for Networked Medical Research (TMF)* provide advice and assistance in the planning, implementation and analysis of epidemiological studies and registries.

6 Template and examples

For the formal description of the Data Dictionary, a Template as Excel file is provided for download:

https://mosaic-greifswald.de/fileadmin/Produkte/Leitfaden_DataDictionary/Vorlage2_v2_en.xlsx

Please use the template as a working basis and expand it as required.

Please send comments, supplements and questions via e-mail to mosaic-project@uni-greifswald.de.

[4] Please pay attention to the Guideline for developing electronic Case Report Forms (eCRFs) provided by the MOSAIC-Project.

7 References

1. German Society of Epidemiology (DGEpi). Guidelines and Recommendations to Assure Good Epidemiologic Practice (GEP). [Online].; 2008 [cited 2014 2 24]. Available from: http://dgepi.de/fileadmin/pdf/GEP_LL_english_f.pdf.

2. Müller D, Augustin M, Banik N, et al. Memorandum Register für die Versorgungsforschung. Gesundheitswesen 2010. 2010 72: p. 824-839.

3. Jensen U. Guidelines for the management of research data (Leitlinien zum Management von Forschungsdaten). Technical Report. Köln: Gesis - Leibniz Institute for Social Sciences, Social Sciences; 2012 Jul.

4. MySQL-Homepage (Schema Object Names). [Online].; 2015 [cited 2015 7 16]. Available from: https://dev.mysql.com/doc/refman/5.0/en/identifiers.html.

5. FHIR Project. Value Set for Codes in HL7 Null-flavors. [Online].; 2014 [cited 2015 2 24]. Available from: http://hl7.org/implement/standards/fhir/null-flavor.html.

6. Open Clinica. Null-Value Codierung in Open Clinica. [Online]. [cited 2015 2 24]. Available from: https://community.openclinica.com/OpenClinica/3.0/doc/glossary.html#nullValues.

7. Bureau Internation des Poids et Mesures. SI Brochure: The International System of Units (SI) [8th edition, 2006; updated in 2014]. [Online].; 2014 [cited 2015 8 24]. Available from: http://www.bipm.org/utils/common/pdf/si_brochure_8.pdf.

8. World Health Organization. International Statistical Classifiaction of Diseases and Related Health Problems, 10th Revision. [Online].; 2010 [cited 2015 4 12]. Available from: http://apps.who.int/classifications/icd10/browse/2010/en.

9. Regenstrief Institute, Inc. LOINC.org. [Online].; 2015 [cited 2015 4 12]. Available from: https://loinc.org/.

10. Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE). Nationales Metadata Repository. [Online].; 2015 [cited 2015 4 12]. Available from: https://mdr.imise.uni-leipzig.de/.

11. Boslaugh S. An Intermediate Guide to SPSS Programming: Using Syntax for Data Management. 1st ed.: SAGE Publications; 2004.