

## What is a DataSchema?

1. Objectives .....	2
2. Context.....	2
3. Structure and Content .....	2
4. Variable Selection .....	3
References.....	3
Figure 1: The 'Household Status' Domain in the Hierarchical Structure of a DataSchema.....	4

*This document is part of a series describing P<sup>3</sup>G's DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) project. For more information on this project, please refer to other protocols and electronic documentation available at: [www.datashaper.org](http://www.datashaper.org)*

## **1. Objectives**

This document presents the general objectives and structure of a DataSchema. It is intended to be used by DataSchema developers and users.

## **2. Context**

The DataSchema lies at the heart of the DataSHaPER project. A DataSchema aims to identify and describe a thematic set of core variables that are of particular value in a specified scientific setting, and which are recommended to optimize potential data sharing between biobanks. DataSchemas can be developed to provide a template for the prospective harmonization of emerging biobanks. They can also be developed with the goal of facilitating retrospective harmonization of existing biobanks. In other words, DataSchemas can facilitate data sharing for an emerging group of studies which will work together on a specific topic, or to provide a tool for existing studies which intend to share data. A growing number of DataSchemas are being created to serve a range of scientific areas of interest (e.g., broad-based epidemiological research, cancer, renal diseases, obesity).

## **3. Structure and Content**

DataSchemas are structured as nested hierarchies. In each DataSchema, the *variables* (i.e., primary units of interest for a statistical analysis) are grouped into *domains* of interest representing, for example, risk factors and outcomes. These *domains* are, in turn, grouped under *themes* representing general areas of interest. Lastly, the *themes* are classified into *modules*, which are groupings based on assessment modes or the type of element measured or collected (e.g., Health and Risk Factor Questionnaires, Physical and Cognitive Measures). Figure 1 illustrates this hierarchical structure.

Each DataSchema is comprised of variables that may be derived from a number of different data sources, including: interview administration; health and risk factor questionnaires; physical and cognitive measures; medical files; sample collection/handling/processing/banking; biochemical measures; registries (e.g., death,

hospitalization, environmental). Variables may be of primary scientific interest in their own right or serve as qualifying factors that contribute to the interpretation of a given variable. A variable may also be complete in itself (e.g., 'Current Cigarette Smoker' [yes/no], 'Weight') or it may be derived from one or several other variables (e.g., the 'Body Mass Index' variable is derived from 'Height' and 'Weight' variables).

Each DataSchema includes:

- A list of variables, each with an associated description, definition and format. Where possible, DataSchema variables are defined so that they can be reliably constructed from standard questionnaires and classifications that are used widely (e.g., IPAQ for physical activity, ISCED for education).
- A list of domains, each with an associated definition; scientific relevance, associations with other risk factors and outcomes, scientific references; overall justifications for their inclusion in the DataSchema; links to relevant ontologies; access to reference questions and questionnaires; and indices and operating procedures that have been selected or developed.
- A list of themes and modules with their respective definitions.

#### 4. Variable Selection

Selection of DataSchema variables is conducted by an interdisciplinary working group and guided by iterative review and consensus methodologies. The variable selection process is described in the *DataSchema Development: Variable Selection* document, and is available online at <http://www.datashaper.org>.

#### References

Fortier *et al.* (2009). Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Submitted*

The International Physical Activity Questionnaire (IPAQ), 2005.  
Available at: <http://www.ipaq.ki.se/ipaq.htm>

International Standard Classification of Education (ISCED), 1997. Available at: [http://www.unesco.org/education/information/nfsunesco/doc/isced\\_1997.htm](http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm)

**Figure 1: The 'Household Status' Domain in the Hierarchical Structure of a DataSchema**

