Problems of Comparability in the German Microcensus over Time and the New DDI Version 3.0

by Jeanette Bohr, Andrea Janßen & Joachim Wackerow
*

Introduction
The Data Documentation Initiative (DDI) was created as an international documentation specification to improve the access to and the analysis of social science data. The DDI can be seen as a reaction to a growing need for data documentation standards, brought about by the increased diffusion of quantitative data in the social sciences, because "[...] accurate use of data depends on access to comprehensive, accurate documentation" (Blank/Rasmussen 2004, 310). Initiated by the Inter-University Consortium for Political and Social Research (ICPSR), the ongoing development of the standard is carried out by the membership-based DDI Alliance. The Alliance is developing the DDI specification, which is written in XML.1 At present, a new version of DDI, whose new possibilities have been discussed at the 2006 IASSIST conference2, is under review. As a contribution to this discussion, the paper will point out an example of use for the new DDI Version 3.0, with emphasis on the new grouping model. The German Microdata Lab at the Centre for Survey Research and Methodology in Mannheim prepares documentation for official data for the scientific community. The documentation of the German Microcensus refers to the DDI standard Version 2.1. As an annually-repeated survey, the German Microcensus contains some changes over time that must be documented. However, one of the biggest limitations of the second version of DDI is the lack of the capacity for documenting repeated surveys. The DDI 2.1 documentation pertains only to a single survey, without an option for recording changes over time in repeated surveys.

The new DDI Version 3.0 and the new grouping structure enables the documentation of comparability of variables over time for repeated studies. This paper explains this structure and gives an example for using the German Microcensus for the application of this new grouping model.

DDI Version 3.0
The major change of DDI Version 3.0 from preceding versions is the increased scope of metadata which can be captured. With DDI 3.0, it is now possible to describe all aspects of the data life cycle (see Thomas 2006a, Nelson 2006). The consideration of the entire data life cycle necessitates a means for describing groups of studies as well as the relationships within collections of comparable studies. The new grouping functionality can be used to define a set of comparable studies which can be described in one single instance. The possibility of documenting information about the comparability of several studies represents a major improvement over previous versions of DDI. Because the new version is still under review, the new possibilities of documentation with DDI 3.0 have yet to be fully explored. This paper will serve as a contribution to that exploration.

The Example: The German Microcensus
The German Microcensus is a representative annual population sample containing structural population data of one percent of all households in Germany. From every survey, a sub-sample of seventy percent is drawn, which is made available to the scientific community as Scientific Use Files. Because of the annual repetition, the broad scope of topics and the large number of interviewees, the scientific use files are suitable for the analysis of social structure and the observation of social change in society. To observe social change with the Microcensus, however, it is necessary to have information about the comparability of variables among census years.

The contents of the Microcensus and the questionnaires are regulated through a law called the "Mikrozensusgesetz." Changes over time in the Mikrozensusgesetz have complicated the comparability of variables over time. Between 1995 and 1996 there was an especially significant change concerning the questionnaire, leading to a variety of changes on different levels. Many of the variables of the 1995 census were split into two variables in subsequent years. In addition to this, there were changes in the question routes of the questionnaires, and in many cases the question texts were changed as well. Further inconsistencies concern the values of the variables and the value labels. All the information about inconsistency among census years must be adequately documented and administered. For this purpose, the new DDI Version 3.0 offers new possibilities.

DDI 3.0: Grouping Model
Within the DDI Version 3.0, variable inconsistency can

be described within the grouping model. It offers the possibility to define any information as a standard on a top level and to capture variations or additions on a lower level. Consequently, the application of the new model permits the documentation of coherences and variations among different census years.

The main improvement the new version offers is the facility for the inheritance of the common characteristics of studies down the hierarchy tree of the metadata. This means a simplification of DDI instances because the common metadata has to be stated only once at the upper level of the grouping structure.

The grouping structure consists of several hierarchical levels:

- The Group as the top level contains common metadata which is inherited down the hierarchy of the grouping structure.

- Subgroups can be created on one or more lower levels. In the following example the subgroups contain the information for a period of predominantly consistent census years.

- The Study Unit represents a single study on which all the lower-level modules depend.

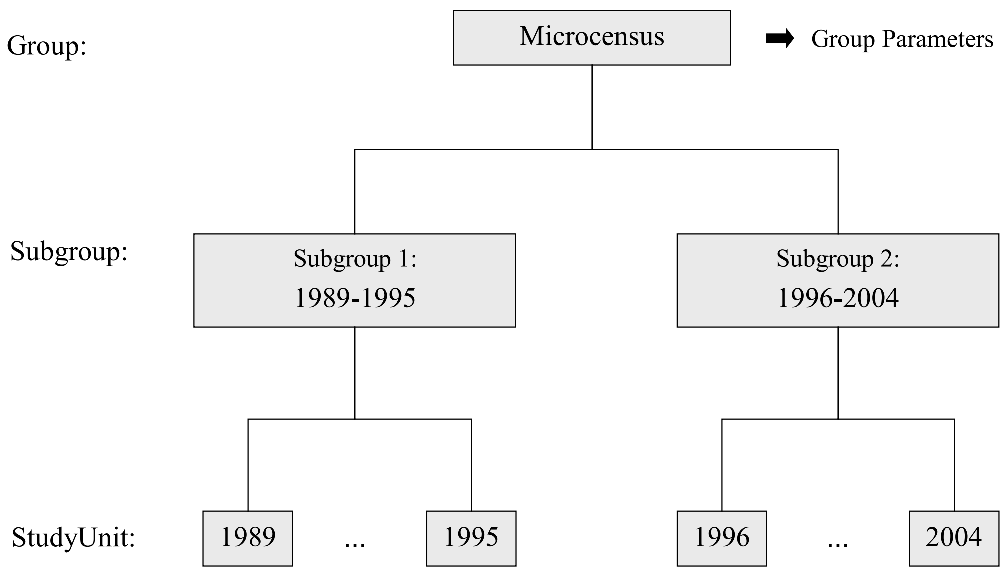
Groups and study units both contain a cluster of modules which describe a study. These modules are: the “Concept” which contains mainly the study description; the “DataCollection” which includes information about the methodology and the instrument; the “LogicalProduct” with most of the material in the data description, and finally “PhysicalDataProduct” and “PhysicalDataInstance” which capture the file description (see Thomas 2006a).

The concept of the inheritance means that classes of specific information always and at any level inherit from their ancestor classes. The specified metadata on the top position of the hierarchy is valid for all studies in this group, unless it is overridden at a lower level. If a piece of information is not valid for a member of the group, the mechanism of local overrides allows for an appropriate replacement of the information on a lower level.

The German Microcensus in the Grouping Model Structure[3]

The organization of the census years within the grouping structure requires a precise knowledge of the data. In particular, the definition of subgroups presupposes a thorough familiarity with all consistencies and changes over time. For the following example, the census years from 1989 to 2004 were taken into consideration.

Figure 1: Hierarchical structure of the German Microcensus in the grouping model



On the top level, the German Microcensus is defined as the Group. This class includes the common metadata, which is shared by all census years such as the general part of the study description or the basic conception. On the Group level, the group parameters have to be defined. This is a set of required properties of the survey, which determines the relationship between the study units of a group (see Thomas 2006b). The group parameters of the German Microcensus are marked in the following table:

Table 1: Group Parameters

Parameter	Tag	Description
TIME	T0	no formal relationship
	T1	single occurrence
	T2	multiple occurrence: regular occurrence: continuing
	T3	multiple occurrence: regular occurrence: limited time
	T4	multiple occurrence: irregular occurrence: continuing
	T5	multiple occurrence: irregular occurrence: limited time
INSTRUMENT	I0	no formal relationship
	I1	Single
	I2	multiple: integrated set of 2 or more instruments used for different subgroups
	I3	multiple: base with topical changes
PANEL	P0	no formal relationship
	P1	single panel surveyed multiple times
	P2	single panel surveyed once
	P3	rolling panel (multiple interviews limited duration)
	P4	different panel each survey
GEOGRAPHY	G0	no formal relationship
	G1	single geography surveyed multiple times
	G2	single geography surveyed once
	G3	rolling geography (multiple interviews limited duration)
	G4	different geography each survey
DATA SETS	D0	no formal relationship
	D1	single data file from a data collection
	D2	multiple data products from a single data collection
	D3	integration of multiple data sets into a single integrated structure
	D4	multiple data files each from a different data collection

Because of changing legal regulations and a significant break in the survey program, two different periods can be defined: the first period including the years from 1989 to 1995 and the second including the years from 1996 to 2004. An extensive variable consistency exists within these periods. Consequently, it proves to be useful to define these two periods as subgroups on a lower level, where most of the variable description is included.

As an alternative, one could define the census years from 1996 on as a standard in the top level and the period before as a subgroup. The disadvantage of this approach is, that the structure is less flexible for accommodating any future significant changes.

Finally, the study units represent the single census years of the German Microcensus.

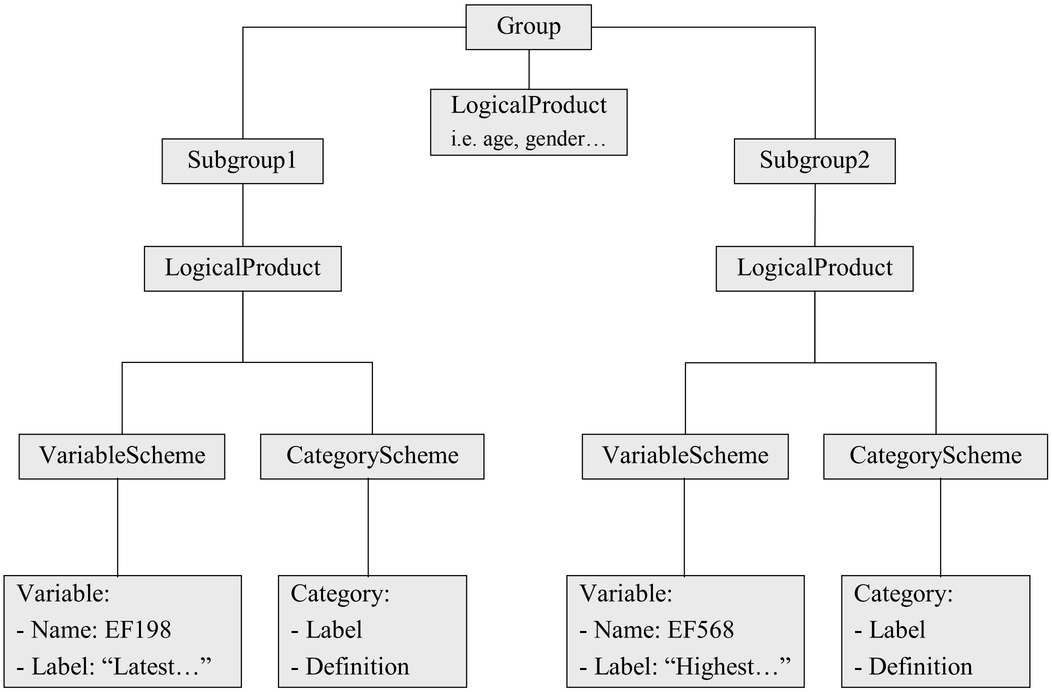
Example 1: Variable Inconsistency between Subgroups
The first example illustrates the possibilities of documenting variable inconsistency between two subgroups. In both periods a variable exists with information about the vocational training certificate for the person who is head of the household. Apart from the different variable names and labels, the contents of the variables differ as well: the variable in Subgroup 1 contains the latest certificate, while the variable in Subgroup 2 contains the highest:

Subgroup 1: 1989-1995	“Latest vocational certificate: head of household” (EF198)
Subgroup 2: 1996-2004	“Highest vocational certificate: head of household” (EF568)

Because of the different meaning of the two variables, we define their variable-specific properties within the specific subgroup. As a result, the information is shared by all members of the subgroup. The relevant metadata for each subgroup in this case are the variable name and label, the category information and the associated question text.

In DDI 3.0, information about the variable and the categories of the variable both have to be defined in the module “LogicalProduct.” The variable attributes “name” and “label” are described in the container or

Figure 2: DDI Structure Example 1: LogicalProduct

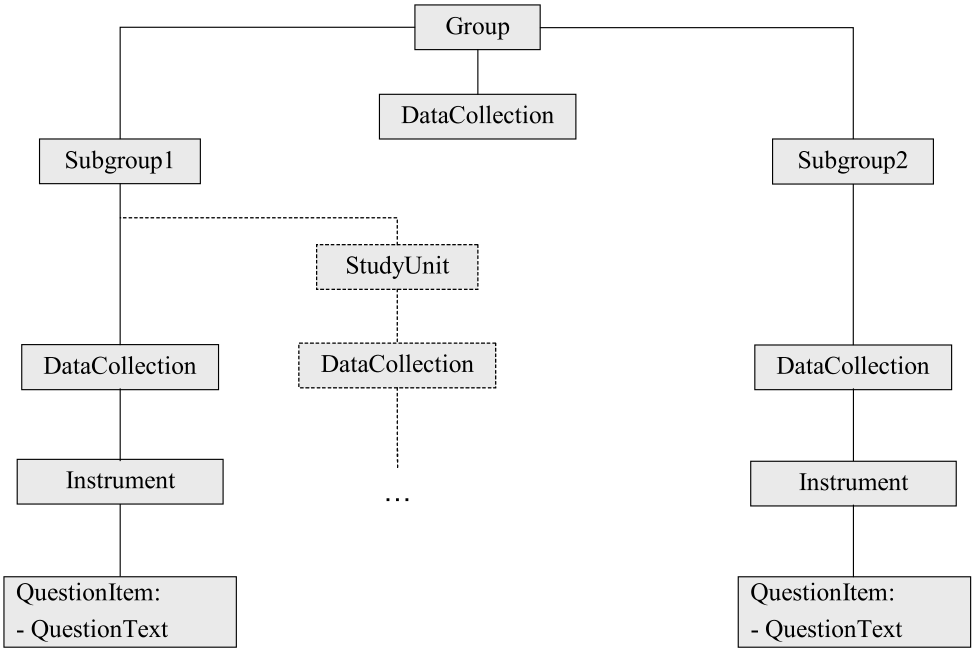


"VariableScheme." The variable categories are described in the container "CategoryScheme" which is illustrated in simplified form in Figure 3. Hence, in both elements the adequate attributes can be described for each subgroup. Variables which are identical for all census years (e.g. age gender) are described on the top-level in the group and are valid for all study units.

The question text for the variables is defined in the module "DataCollection" on the subgroup level. It is described in

Subgroup 2: 1996-2004	**Study Unit: 2003**
"Type of attended school" (EF72) 1 Class 1 to 4 2 Class 5 to 10 3 Class 11 to 13 (sixth form) 4 Vocational School 5 University of applied Sciences 6 University 9 Non Response Not Applicable	"Type of attended school" (EF74) 1 Class 1 to 4 2 Class 5 to 10 3 Class 11 to 13 (sixth form) 4 Vocational School 5 Vocational Preparatory School 6 Vocational School with middle Graduation 7 Vocational School with higher Graduation 8 Technical College, University of Cooperative Education 9 Graduation of College and Advanced Administrative Studies 10 University of applied Sciences 11 University 12 PhD Program 99 Non Response Not Applicable

Figure 3: DDI Structure Example 1: DataCollection



the element "QuestionText" of the container "Instrument." Besides the question text, there is specific information about the questionnaire such as the question number which changes from year to year. This annually-changing information should be defined in the single study units on a lower level. This example illustrates how important the knowledge of the consistencies and changes among the census years is for the correct definition of the information in the hierarchical structure of DDI.

Example 2: Variable Inconsistency within Subgroups
The next example deals with a change on the category level within one subgroup. In Subgroup 2 (1996 to 2004) the variable "Type of attended school" (EF72) contains seven categories; in 2003 the variable differentiates among more categories regarding vocational school and university. In addition, the variable name has changed to EF74.

Because of this variation within the subgroup, the shared information about the variable categories on the subgroup level has to be overridden for the specific year. This can be realized by defining the information about the categories at the study unit level for the census year 2003. Furthermore, the variable name and the response categories of the question should be defined at the study unit level, because these elements are different from the description in the subgroup, too.

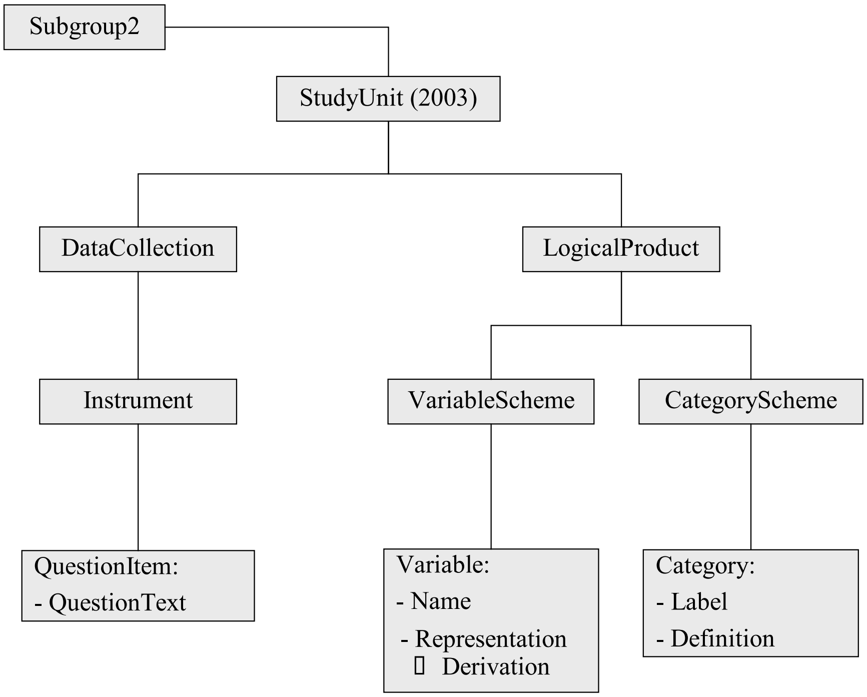
For the description of the categories and the variable name, we need the module "LogicalProduct" again. But now we have to use the module on the study unit level. However, the variable label does not need to be specified in the study unit because this information is inherited from the subgroup.

As in the previous example, the question text will be stated in the module "DataCollection." If the same variation of the variable exists over several years, for instance in 2003 and the following years, the definition of a new subgroup for this information below Subgroup 2 could be useful.

Furthermore, DDI gives the opportunity to document the information about the comparability of non-identical variables. In case of the given example, a new variable can be created by combining the categories of the variable in 2003, which is comparable to the standard variable in the subgroup. Information about the new variable as well as the needed recode job can be stated in the element "Derivation" of the container "VariableScheme."

The possibility of documenting aspects of comparability marks an important advantage in terms of the documentation of the German Microcensus. Due to frequent variations on the category level between several years, many variables are comparable but not identical. The possible creation of explicit comparability is important

Figure 4: DDI Structure Example 2



information for using the data for analyses over time.

Conclusion
All in all, it can be stated that the innovations of the new DDI version will improve the data documentation regarding comparability over time. For the German Microcensus as an annually repeated survey with some breaks in the survey program, the grouping structure will simplify the documentation over time. For a group of studies, the migration from Version 2.1 to 3.0 may become quite complex, but the possibilities for grouping multiple studies and for overriding information on a lower level offer a more efficient means of documentation, especially in the case of variable inconsistency between periods of time or single years.

However, there are some limitations of the grouping model. The efficient use of the grouping mechanism for documenting comparability over time requires exact knowledge of the documented data and needs a lot of preliminary work. Every extension of the grouping structure results in a higher branching as well, so that the administration of the DDI structure may become quite complex. Moreover the flexibility of the structure is limited once a standard has been defined. As a consequence, the basic structure can not be modified to accommodate potential future changes in the survey program. This limitation concerns not only DDI, but is inherent in working with comparative standards.

Nevertheless, overall DDI 3.0 provides an instrument for a better managing and processing of metadata. These possibilities should be used to ensure high quality data documentation which is comparable in an international context.

* Jeanette Bohr, Andrea Janßen und Joachim Wackerow, Centre for Survey, Research and Methodology (ZUMA), Mannheim, Centre for Survey Research and Methodology, ZUMA, P.O. Box 12 21 55 - 68072 Mannheim - Germany. Contact: bohr@zuma-mannheim.de. http://www.gesis.org/en/social_monitoring/GML/index.htm.

References

Blank, Grant and Karsten Boye Rasmussen (2004). The Data Documentation Initiative. The Value and Significance of a Worldwide Standard. In: Social Science Computer Review, Vol. 22, No. 3, 307-318.

Nelson, Chris (2006). DDI 3.0. Conceptual Model. IASSIST Conference 2006 Presentation. http://www.iassistdata.org/conferences/2006/presentations/W5_DDI.ppt

Thomas, Wendy (2006a). Codebook Centric to Life-Cycle

Centric: In the beginning .... IASSIST Conference 2006 Presentation. http://www.iassistdata.org/conferences/2006/presentations/W5_Thomas_InTheBeginning.ppt

Thomas, Wendy (2006b). Organizing Groups. IASSIST Conference 2006 Presentation. http://www.iassistdata.org/conferences/2006/presentations/W5_Thomas_Groups.ppt

Endnotes

[1] http://www.icpsr.umich.edu/DDI/org/index.html

[2] http://www.iassistdata.org/conferences/2006/presentations/

[3] In the following, the DDI structure examples are illustrated graphically, not in XML.