Microdata Information System - MISSY

*by Andrea Janssen and Jeanette Bohr**

Introduction

In recent years, the number of official microdata sets accessible as Scientific Use Files has increased significantly in Germany. These microdata are of great interest to both economists and social scientists but are not, however, easy to work with.

Official microdata are surveyed to meet the data requirements of the German Federal Statistical Office. The data contain special classification types which need to be documented for the user. Users from the scientific community require more than superficial descriptions of a particular dataset; they also require detailed information pertaining to every variable in it.

Such an example of a German information system designed to fulfill the need of researchers, is the Microdata Information System (MISSY) presented below1. MISSY contains metadata or "data about data", (Jacobs 2006) about the German Microcensus. MISSY piloted a project containing the descriptions for two census years 1995 and 1997.

The next section introduces the Microcensus and describes the functions and benefits of MISSY. The last section introduces the steps necessary to fully implement the system. Questions concerning general rules for documenting and presenting metadata derived from the experiences in the first phase of the project are also addressed.

The German Microcensus

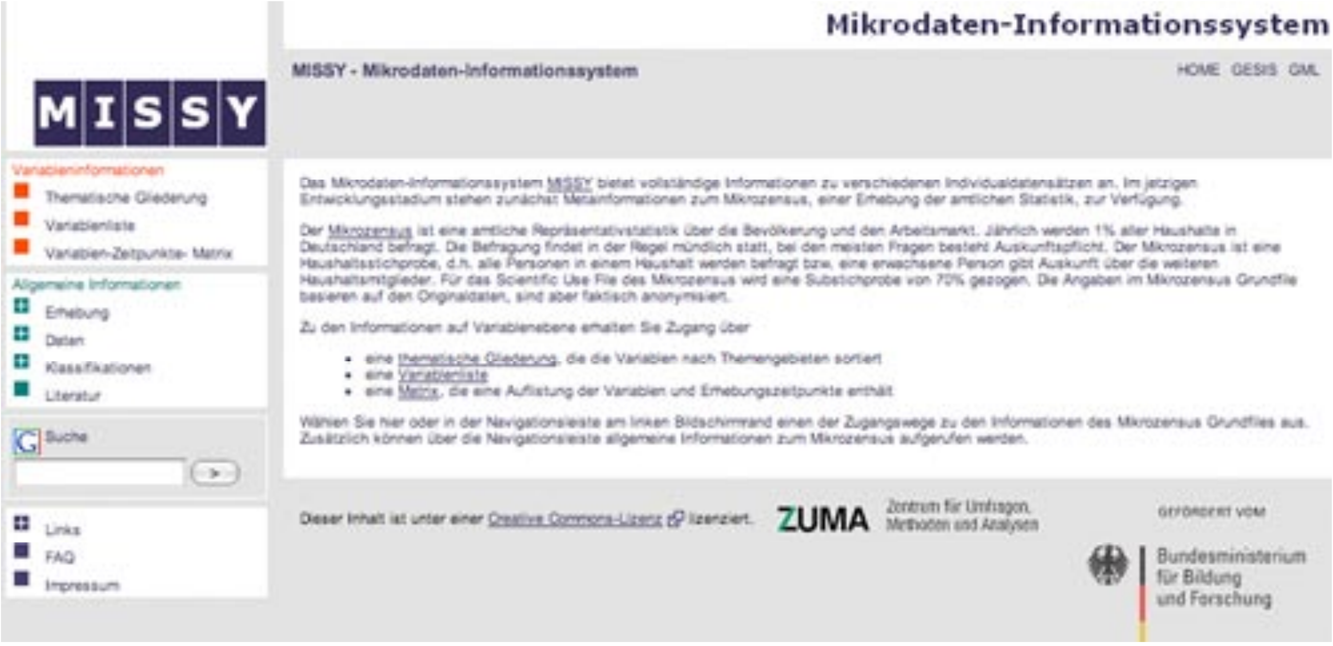
The Microcensus is the biggest continuing survey in Germany. Conducted annually since 1957 by the Federal Statistical Office, it samples one percent of all German households or approximately 820.000 people. The main topics of the Microcensus are occupation and qualification, labor markets and household and family structures. Every four years the Microcensus contains additional questions about health or housing conditions, for example. The large sample size and the broad scope of topics make the Microcensus an invaluable data source for different scientific questions of varying complexity. For example, the Microcensus enables one to examine higher education among relatively small groups of immigrants, e.g. Italians or Greeks.

The Microcensus cannot be accessed by the scientific community in its entirety, but the Federal Statistical Office extracts a 70 per cent subset and provides it to researchers.

The Microcensus has attributes that are not commonly included in social sciences surveys: the classifications of professions (KldB – Klassifikation der Berufe) and economic sectors (WZ – Wirtschaftszweige) are used only in the official statistics and require some explanation to the researcher. Another characteristic of the Microcensus is its vast quantity of derived variables, the so-called "Bandsatzerweiterungen und Typisierungen", whereby the latter are based on different concepts of families and living arrangements. The generation of these variables is not easy to comprehend; again they are only partially accessible to the scientific community. As a result, to work competently and efficiently with the Microcensus data, the researcher requires information exceeding what a superficial description of the dataset can provide. For this reason MISSY was developed.

MISSY

MISSY is a product of the German Microdata Lab (GML), formerly named the Department for Microdata at the ZUMA (Centre for Survey research and Methodology) in Mannheim.2 Since the 1980s, the GML's focus has been on the Microcensus and as part of this focus it has offered an array of comprehensive services to support use of the Files. The GML, in collaboration with the Federal Statistical Office, ensures that the procedures necessary for anonymizing the data to protect confidentiality are in place. All Files are checked prior to being released to the scientific community and comprehensive documentation of the Files is created. As well, the GML provides support to researchers by offering advice on both the methodology and the content. To facilitate work with, e.g., classifications unique to the Microcensus, microdata tools are developed. Finally, and of equal importance, the GML organizes user conferences and workshops to promote the advantages of the Microcensus data for scientific research and to enable and increase the opportunity for scientific communication among researchers (Lüttinger et al. 2004).



MISSY - Mikrodaten-Informationssystem

Das Mikrodaten-Informationssystem MISSY bietet vollständige Informationen zu verschiedenen Individualdatensätzen an. Im jetzigen Entwicklungsstadium stehen zunächst Metainformationen zum Mikrozensus, einer Erhebung der amtlichen Statistik, zur Verfügung.

Der Mikrozensus ist eine amtliche Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt. Jährlich werden 1% aller Haushalte in Deutschland befragt. Die Befragung findet in der Regel mündlich statt, bei den meisten Fragen besteht Auskunftspflicht. Der Mikrozensus ist eine Haushaltsstichprobe, d.h. alle Personen in einem Haushalt werden befragt bzw. eine erwachsene Person gibt Auskunft über die weiteren Haushaltsmitglieder. Für das Scientific Use File des Mikrozensus wird eine Substichprobe von 70% gezogen. Die Angaben im Mikrozensus Grundfile basieren auf den Originaldaten, sind aber faktisch anonymisiert.

Zu den Informationen auf Variablenebene erhalten Sie Zugang über

- eine thematische Gliederung, die die Variablen nach Themengebieten sortiert
- eine Variablenliste
- eine Matrix, die eine Auflistung der Variablen und Erhebungszeitpunkte enthält

Wählen Sie hier oder in der Navigationsleiste am linken Bildschirmrand einen der Zugangswege zu den Informationen des Mikrozensus Grundfiles aus. Zusätzlich können über die Navigationsleiste allgemeine Informationen zum Mikrozensus aufgerufen werden.

Dieser Inhalt ist unter einer Creative Commons-Lizenz lizenziert.

Figure I: MISSY

MISSY facilitates research based on the German Microcensus. Gathering all the necessary metadata incorporating the knowledge of the GML is the first step; this includes official documents of the Federal Statistical Office.The second step is one that connects all the metadata in a way that considers the textual relationships between the data and the enquiries of social scientists and economists. The implementation accomplished by MISSY is based on the DDI (Data Documentation Initiative) 2.1 standard.

MISSY is an exclusively German system; there are two reasons for it being unilingual. At first, researchers are forbidden from using the data abroad. The second and more important reason is that all documents and descriptions of the Microcensus are in German making a knowledge of the German language essential. For demonstration purposes the most important expressions in the following examples have been translated.

To classify the metadata type, MISSY utilizes the categories of Sundgren's dimensions. The metadata for the Microcensus includes both pragmatic and semantic as well as syntactic aspects (Fischer 2005, Sundgren 2003). This means that MISSY encompasses data answering questions of why (pragmatic aspects), what (semantic aspects) and how (syntactic aspects). In the Microcensus documentation, it is helpful to differentiate between general information about the entire study and specific information about the variables. The difference between the elements "study description" and "data description" is found as defined in accordance with the "Data Documentation Initiative" (Jacobs and Thomas 2006).

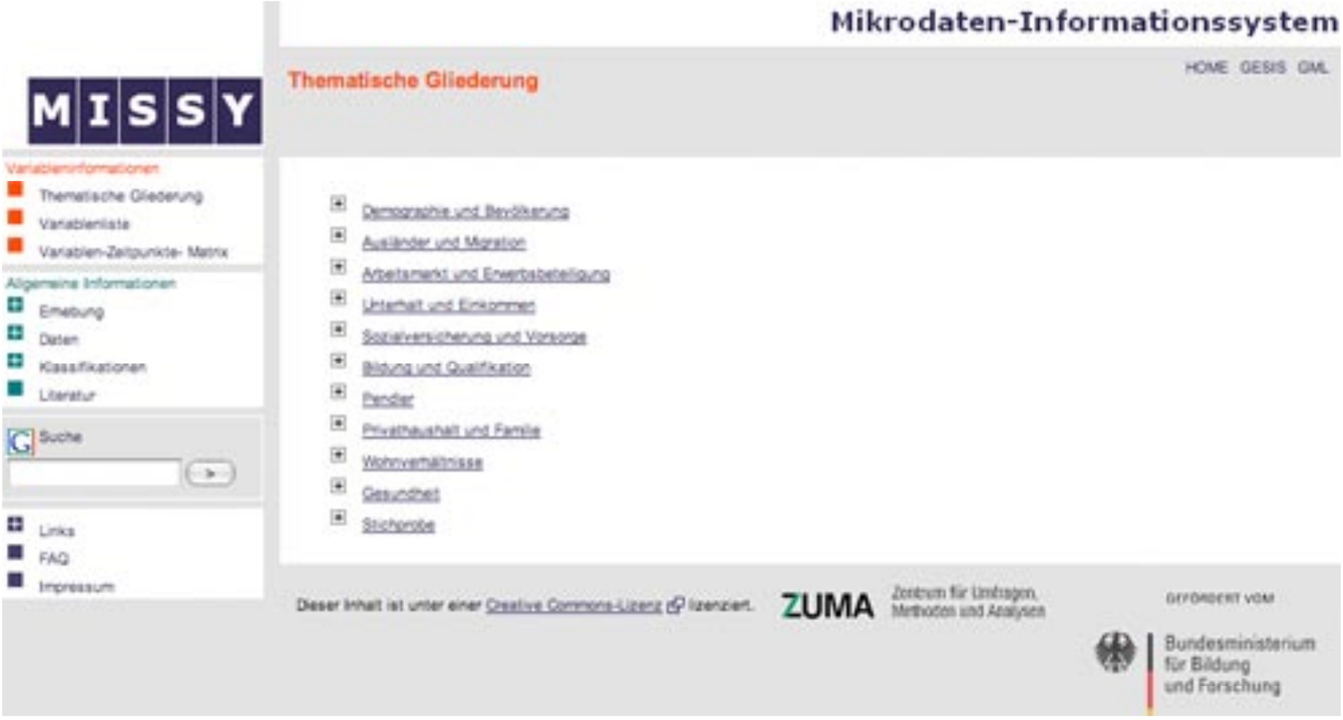
Note that in the middle of the screen there are multiple access points for retrieving specific information. Furthermore, there is a brief overview of the function of MISSY and there are also links to more information about the Microcensus and MISSY. In addition to this "main entrance" for access to specific information, the short list on the left sidebar of the screen under the red header "Variableninformationen" (specific information) can be used as well. The second list with the green header "Allgemeine Informationen" contains general information only about the Microcensus: an introduction, questionnaires, codebooks, interviewer guides, frequencies and some tips and recommendations on working with the data. Information about classifications used by the Federal Statistical Office or the scientific community is included here. For an easier navigation of the MISSY pages all specific information has red headers and all general information has green headers.

Points of Access

The different points of access were designed to simplify the search for variables while recognizing the varying needs and skills of the users. The first point of access is a list of all variables, subdivided by census year. This is useful when information about a specific variable for a particular year is required and it is preferable that the user already has some knowledge of the structure of the Microcensus.

An easier way of access, albeit longer, is given by the thematic structure illustrated below:

Thematic access is appropriate when the researcher is



Dieser Inhalt ist unter einer Creative Commons-Lizenz lizenziert.

Figure II: Thematic Structure

interested in a specific subject or field of research and wants to know if the Microcensus has relevant content. To start with, the researcher may choose from eleven topics leading to the secondary level; at this point there are two links. The first is to publications based upon the Microcensus that pertain to specific subjects, “ethnic minorities and migration” being one example (see fig. II). This makes it easy for the investigator to determine what research might be undertaken or see what research has already been done based on the Microcensus. The second link connects to tables containing examples of analyses. Again, using “ethnic minorities and migration” as an example, the user will find multiple tables including one which shows a comparison of the graduation rates between the German and Turkish population. The tables were created to assist novice data users, e.g. students. The aim is to encourage researchers and future researchers to use official microdata in their analysis.

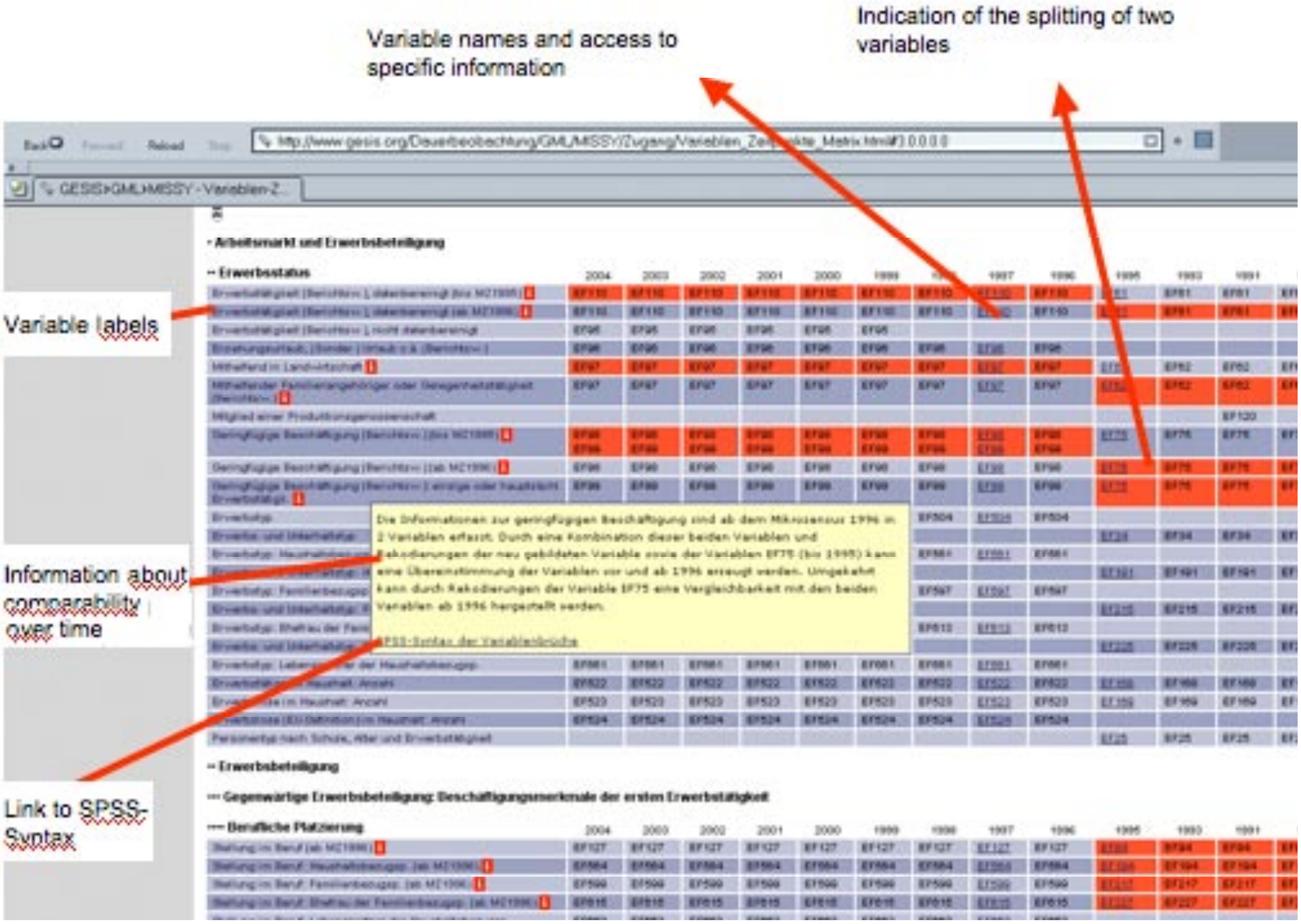
The fastest method for obtaining specific information is via a matrix containing all variables for every year covered by the Scientific Use Files (see fig. III). The variables names in the matrix cells are linked to specific information about the variables. Furthermore, the matrix provides an overview of characteristics surveyed in specific years.

Because the Microcensus is conducted annually, it can be used to address questions requiring a consistent long-term view to observe social change in society. In order to examine longer time periods the researcher requires information about the comparability of the variables over a specific timeframe. The matrix presents the most important changes that have occurred in the variables.

There was a significant change to the Microcensus questionnaire in 1996; many variables were split into two, marked in the matrix. The changes are indicated and explained in the tool tips. When it is possible to generate comparisons of variables, links to SPSS-Syntax are provided. With these instructions, comparisons between many of the variables, both before and after 1996, can easily be made.

Specific information: variables documentation

For each variable, all available metadata are centralized on a single site. Not only are the variable labels included, but also the text of the questions and related notations, if available. There is information about the guiding filters of the questionnaire or what attributes the respondent had to fulfill in order to be asked this special question. The value labels and frequencies give first impressions of the variable’s distribution. In the first lines of the variable description are links to shortcuts to detailed information of comparable variables for other years and for different levels of the thematic structure. These links are marked with red buttons to create visual consistency with the list containing the different possibilities of access to specific information. Analogously, the links to general information that could be


Variable names and access to specific information
Indication of the splitting of two variables
Variable labels
Information about comparability over time
Link to SPSS-Syntax
http://www.gesis.org/Dauerbeobachtung/GML/MISSY/Zugang/Variablen_Zeitpunkte_Matrix.html#3.0.0.0.0
GESIS>GML>MISSY - Variablen-Z...

- Arbeitsmarkt und Erwerbsbeteiligung

-- Erwerbsstatus

Die Informationen zur geringfügigen Beschäftigung sind ab dem Mikrozensus 1996 in 2 Variablen erfasst. Durch eine Kombination dieser beiden Variablen und Rekodierungen der neu gebildeten Variable sowie der Variablen EF75 (bis 1995) kann eine Übereinstimmung der Variablen vor und ab 1996 erzeugt werden. Umgekehrt kann durch Rekodierungen der Variable EF75 eine Vergleichbarkeit mit den beiden Variablen ab 1996 hergestellt werden.

SPSS-Syntax der Variablenbrüche

-- Erwerbsbeteiligung

--- Gegenwärtige Erwerbsbeteiligung: Beschäftigungsmerkmale der ersten Erwerbstätigkeit

---- Berufliche Platzierung

Figure III: Matrix

of interest according to the particular variable are marked with green buttons. With these links, the researcher will be directed to the exact reference in the questionnaire, the codebook or the interviewer guide that contain the information concerning the variable of interest.

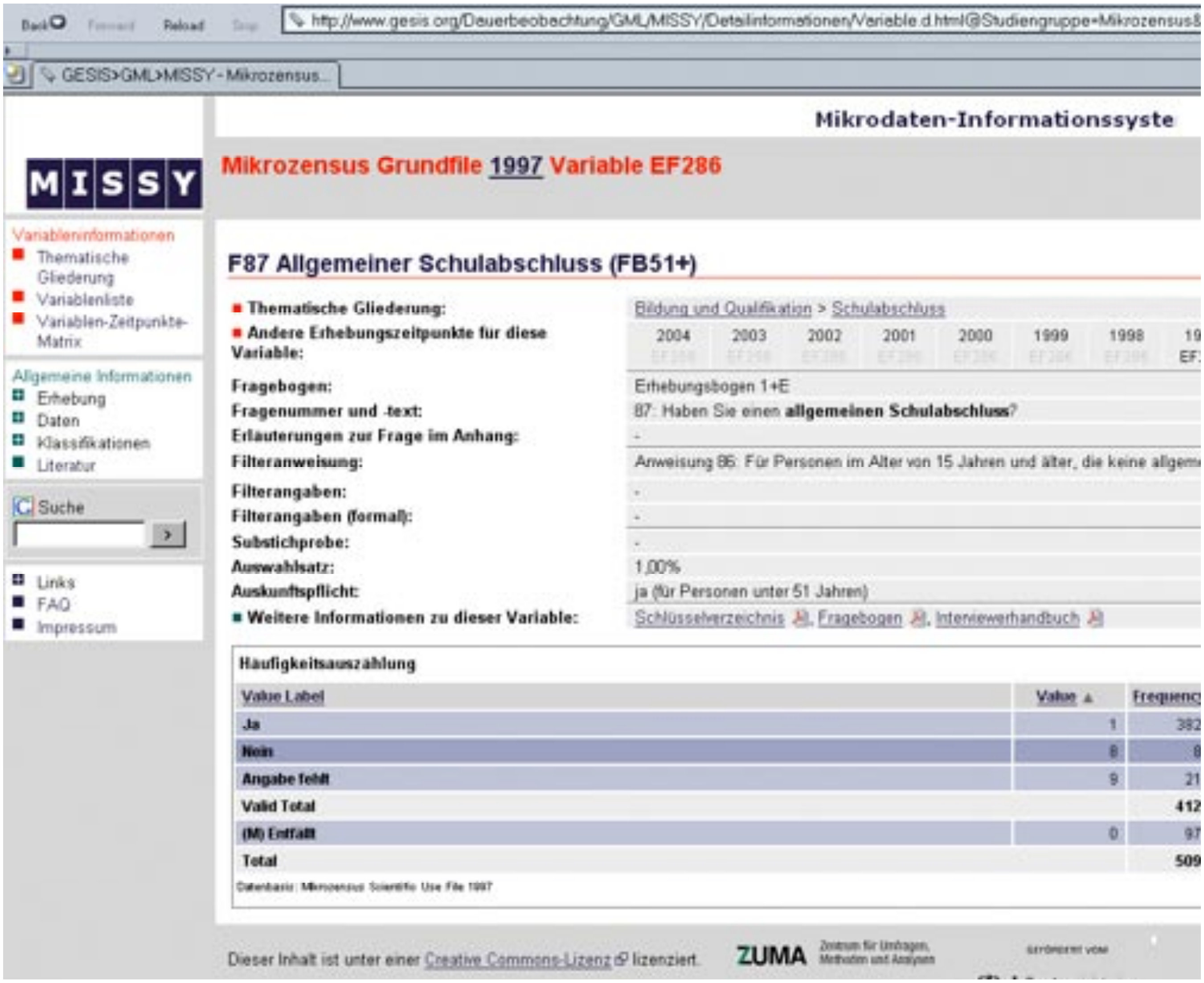
As stated in the introduction, the Microcensus contains a variety of derived variables which cannot be tracked in their composition. Information about the generation of these variables is documented for the internal use of the Federal Statistical Office only and cannot be accessed via the Internet. MISSY provides an additional link from the special information about derived variables that point to a site on which the generation of this particular variable is described. If the generation of the variable is based upon a special concept of families, households or living arrangements used by the Federal Statistical Office, another link to a description of these concepts is provided. Furthermore, researchers can go to a catalogue that contains definitions of the terms used in the Microcensus. Below is an example for a description of the variable "Type of working hours of the reference person of the family":

This information makes it relatively easy to use even rather complicated variables in an appropriate way.

Conclusion

What conclusions can be drawn about the implementation of an information system for microdata? First of all, the concept of the system requires knowledge and research experience with the particular data. The special characteristics of the data should be understood and adequately documented. Secondly, knowing the data should make it possible to connect different kinds of information about the contents and thereby facilitate the search on particular topics. Ideally researchers should find not only all information they are looking for but also other helpful information that they may not even know existed.

Another important point to appreciate after the first


Back
Forward
Reload
Stop
http://www.gesis.org/Dauerbeobachtung/GML/MISSY/Detailinformationen/Variable.d.html@Studiengruppe=Mikrozensus
GESIS>GML>MISSY - Mikrozensus...
Mikrodaten-Informationssyste
MISSY
Variableninformationen
Thematische Gliederung
Variablenliste
Variablen-Zeitpunkte-Matrix
Allgemeine Informationen
Erhebung
Daten
Klassifikationen
Literatur
Suche
Links
FAQ
Impressum

Mikrozensus Grundfile 1997 Variable EF286

F87 Allgemeiner Schulabschluss (FB51+)

Thematische Gliederung:	Bildung und Qualifikation > Schulabschluss
Andere Erhebungszeitpunkte für diese Variable:	2004 2003 2002 2001 2000 1999 1998 19
Fragebogen:	Erhebungsbogen 1+E
Fragenummer und -text:	87: Haben Sie einen **allgemeinen Schulabschluss**?
Erläuterungen zur Frage im Anhang:	-
Filteranweisung:	Anweisung 86: Für Personen im Alter von 15 Jahren und älter, die keine allgem
Filterangaben:	-
Filterangaben (formal):	-
Substichprobe:	-
Auswahlsatz:	1,00%
Auskunftspflicht:	ja (für Personen unter 51 Jahren)
Weitere Informationen zu dieser Variable:	Schlüsselverzeichnis, Fragebogen, Interviewerhandbuch

Häufigkeitsauszählung

Value Label	Value ▲	Frequenc
Ja	1	382
Nein	8	8
Angabe fehlt	9	21
Valid Total		412
(M) Entfällt	0	97
Total		509

Datenbasis: Mikrozensus Scientific Use File 1997

Dieser Inhalt ist unter einer Creative Commons-Lizenz lizenziert.

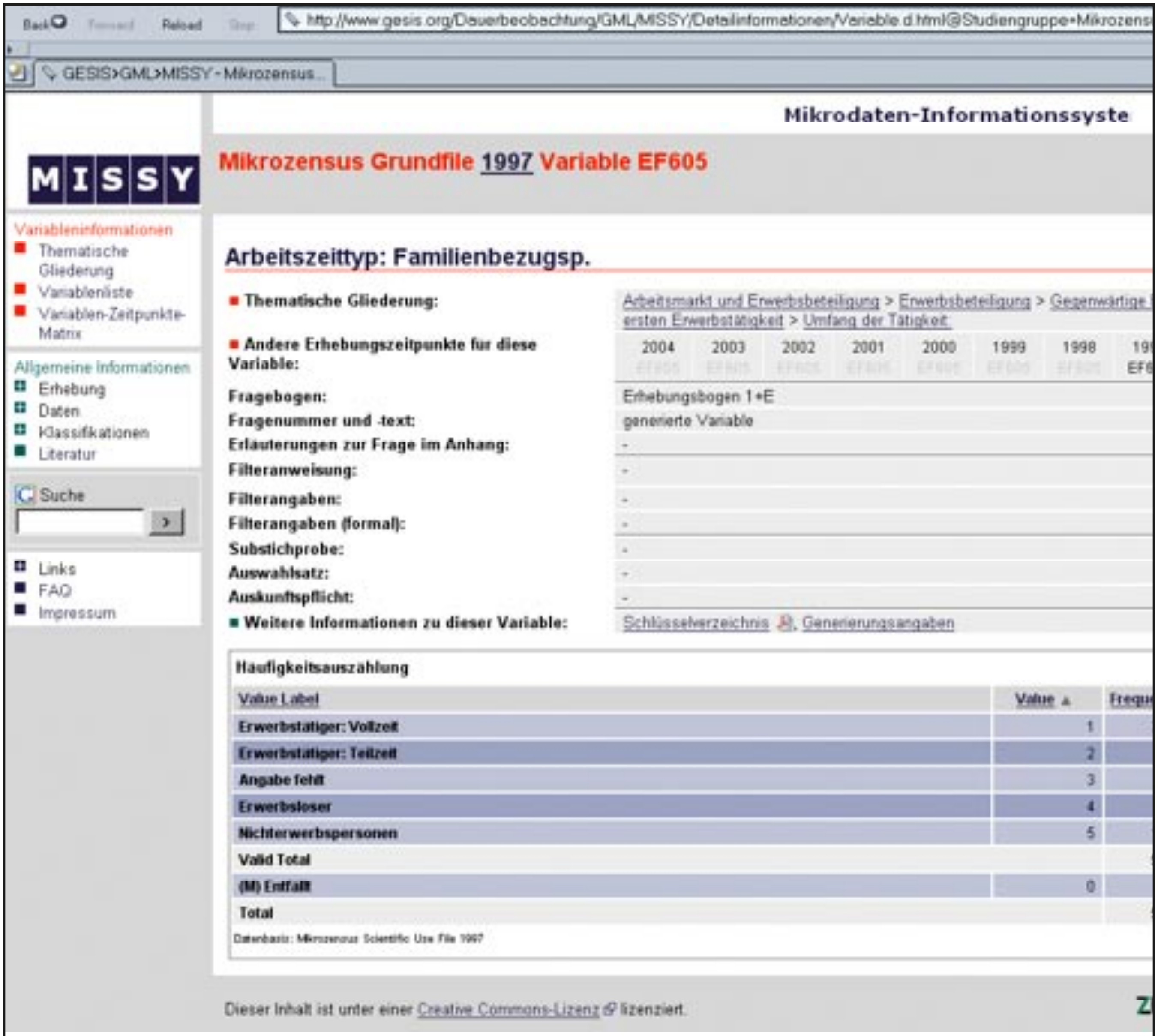
Figure IV: Variable Information

implementation of an information system of course is to ensure a maximum of usability. The next step is to have experts and users of Microcensus data analyze MISSY's performance and make recommendations for improvement. Following this, they are plans to extend MISSY by including all available Microcensus Scientific Use Files. Two more specialized files will be added: the Panel File and the Regional File. Because of the concept of the Microcensus as a rotating panel the Panel File would include four years of census microdata. The Regional File contains microdata of a very differentiated regional level but on a less differentiated topic level to ensure the necessary confidentiality.

To include these new types of data sets in MISSY adequately, some modifications to the information system will be required. Another emphasis will be the extension of the category "tables" that provide an overview of the research possibilities using the Microcensus. An exercise-based introduction into working with Microcensus data is planned. With this concept the main focus of collecting and providing metadata for datasets will be expanded in a direction with implications for more practical and concrete advice for special problems that arise when working with Microcensus data. The result will be that the proportion of metadata with syntactic aspects will increase in MISSY.

References

Fischer, Birgit (2005). Metadaten in der amtlichen Statistik im internationalen Vergleich. Berliner Handreichungen zur Bibliothekswissenschaft, No.45. Berlin: Institut für Bibliothekswissenschaft der Humboldt-Universität zu


Back
Forward
Reload
Stop
http://www.gesis.org/Dauerbeobachtung/GML/MISSY/Detailinformationen/Variable.d.html@Studiengruppe=Mikrozens
GESIS>GML>MISSY - Mikrozensus...
Mikrodaten-Informationssyste
MISSY
Variableninformationen
Thematische Gliederung
Variablenliste
Variablen-Zeitpunkte-Matrix
Allgemeine Informationen
Erhebung
Daten
Klassifikationen
Literatur
Suche
›
Links
FAQ
Impressum

Mikrozensus Grundfile 1997 Variable EF605

Arbeitszeittyp: Familienbezugsp.

■ Thematische Gliederung:	Arbeitsmarkt und Erwerbsbeteiligung > Erwerbsbeteiligung > Gegenwärtige ersten Erwerbstätigkeit > Umfang der Tätigkeit:
■ Andere Erhebungszeitpunkte für diese Variable:	2004 2003 2002 2001 2000 1999 1998 19 EF6
Fragebogen:	Erhebungsbogen 1+E
Fragenummer und -text:	generierte Variable
Erläuterungen zur Frage im Anhang:	-
Filteranweisung:	-
Filterangaben:	-
Filterangaben (formal):	-
Substichprobe:	-
Auswahlsatz:	-
Auskunftspflicht:	-
■ Weitere Informationen zu dieser Variable:	Schlüsselverzeichnis, Generierungsangaben

Haufigkeitsauszählung

Value Label	Value ▲	Frequ
Erwerbstatiger: Vollzeit	1	
Erwerbstatiger: Teilzeit	2	
Angabe fehlt	3	
Erwerbsloser	4	
Nichterwerbspersonen	5	
Valid Total		
(M) Entfallt	0	
Total		

Datenbasis: Mikrozensus Scientific Use File 1997

Dieser Inhalt ist unter einer Creative Commons-Lizenz lizenziert.

Figure V: Explanation of the generation of a variable

Berlin.

Jacobs, Jim (2006). Looking into the future... IASSIST Conference 2006 Workshop Presentation. http://www.iassistdata.org/conferences/2006/presentations

Jacobs, Jim, and Wendy Thomas (2006). Evolution of Data Documentation. IASSIST Conference 2006 Workshop Presentation. http://www.iassistdata.org/conferences/2006/presentations/

Lüttinger, Paul, Bernhard Schimpl-Neimanns, Georg Papastefanou, and Heike Wirth (2004). The German Microdata Lab at ZUMA: Services Provided to the Scientific Community. Schmollers Jahrbuch 124, 455-467

Sundgren, Bo (2003). Developing and implementing statistical metadata systems. http://www.epros.ed.ac.uk/metanet/deliverables/deliverables.html

* Andrea Janssen and Jeanette Bohr. Contact: Andrea Janssen, Centre for Survey Research and Methodology, ZUMA P.O. Box 12 21 55 - 68072 Mannheim - Germany. email janssen@zuma-mannheim.de.

Footnotes

[1] http://www.gesis.org/Dauerbeobachtung/GML/MISSY/

[2] http://www.gesis.org/en/social_monitoring/GML/index.htm