6 Glimpsing into the Blackbox: Data Managing and Cleaning Processes

Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Stephanie Stuck and Sabrina Zuber

6.1 Database management in SHARE – an overview

A data collection effort such as SHARELIFE entails a large amount of work which usually is neither noticed by the respondent nor by the researcher who finally uses the data. It involves several steps around what is called "cleaning" the data – a necessary process when over 1,000 interviewers produce about 28,000 interviews. This chapter is meant to give an overview of the tasks and thus provides some explanation of why such a long time passes from when the data are collected to when they are finally released. While this chapter is in the SHARELIFE methodology volume, it provides an overview of all SHARE database tasks and challenges alike.

As in all parts of SHARE, high standards are applied as well in the data base management concerning cross national comparability and harmonisation, which requires tremendous coordination and cooperation between the different actors involved. The central coordination unit of the SHARE data base management is located at the Mannheim Research Institute for the Economics of Ageing (MEA) in Mannheim, Germany, while the more technical part of data base management is done by CentERdata in Tilburg, the Netherlands. Among other things, CentERdata is responsible for the collection of all data from the survey agencies, provides the internal versions of the raw data and, once a data release has been finalized, distributes the public release data versions on the web site. The teams of researchers and operators in each country also play an important role in data base management: First, they are responsible for all issues that require knowledge of the national languages. Thus, country teams have to check all interviewer remarks and provide programmes to correct data accordingly (also see section 6.3). In addition, they write programmes to correct wrong IDs, erroneous household compositions or demographic information (see also section 6.2). Furthermore, they are involved in developing coding schemes and apply them to the open answers given in the field (also see section 6.4). The imputation group located in Padua and Salerno, Italy, is another important actor for the data base management. They provide multiple imputations of missing data based on a first cleaned version of the data, which already takes into account the corrections mentioned in the previous steps. Similarly, the group working on the survey weights – located in Rome, Italy – uses the first corrected data to compute different weights for all SHARE countries centrally.

To ensure common standards, coordination of the data base management tasks is essential, and thus the different actors meet on a regular basis. Usually these (half to one day) meetings take place together with regular SHARE meeting every three to six months. Here, MEA, CentERdata, the so called country team operators, and sometimes members of the imputations team meet to discuss strategies, solve appearing problems and to agree on a work schedule. The central coordination team provides instructions and programme templates for the tasks locally processed. Country team operators then execute the instructions and write programmes for the respective country. MEA pools and runs all these programmes centrally to produce new versions of the data.

All data base management processes are basically aimed at generating two main products: the public release data and the so-called preload data for the next wave. While the “public release data” are those used by the scientific community to do research, “preload data” are data which come from a previous wave of data collection and are used in the interview of a new wave. Using preload data involves having all demographic information of a previous respondent loaded into the Sample Management System (see Chapter 3), such as information on gender, age, previous interview status, as well as details on household composition. With this information, the interviewer can check the details on the respondent before the interview. During the interview, just checking for changes is quicker, and the interviewer as well as the respondent may feel more comfortable when reiteration of known facts is not necessary.

The first steps of data cleaning are done for both purposes – public release and preload data – in common and then the procedures are split. The first adjustments are sometimes already necessary during the fieldwork on interviewer laptops directly, and are done by CentERdata. After fieldwork, the process of data cleaning starts with corrections according to interviewer remarks (see section 6.3). These are followed by corrections that come from checks of matching between modules within and across waves. The main focus here is on the identification of household members (via their ID numbers) and basic demographic characteristics (see section 6.2). The resulting corrected data base is used for the public release as well as for the basis of the preload data for the next wave. Only afterwards, preload information is combined with information from other sources, e.g. if survey agencies have knowledge on whether a respondent has deceased.

Public releases of SHARE data further require certain changes to the raw data to make the files user-friendly: Main issues are data formatting and the provision of generated variables (e.g. coding of education into ISCED categories). Furthermore, for data protection purposes open answer variables are not included as text, but in their direct form but in category coded form only (see 6.4 and 6.5).

As indicated in figure 6.1, which gives an overview on all data management tasks and the involved actors, database management – especially in a panel study – is not a one-way process, but rather a procedure that includes feedback loops. Some problems in the data can only be detected if new information is available, for example from feedback of interviewers or from new data of a next wave. User questions and hints from their side are also very useful to spot errors. Corrections based on such information are then included in the next version of the data base that is used again for the next release and the next preload data.

Last but not least, it is important to point out that data cleaning in SHARE is done very conservatively. The general philosophy is that respondents are experts of their own lives and that their answers need to be taken seriously and at face value. Modifications of the original data are only made if it is certain that a specific value is wrong and if, beyond that, reliable information on the correct value is accessible (e.g. from an interviewer remark). Data are never changed based on mere plausibility assumptions. However, if implausible values occur, additional indicating variables might be added to the data. The user is free to decide how to handle ambiguous or contradicting information, but is always urged to be as careful as possible with assumptions and changes to the data.

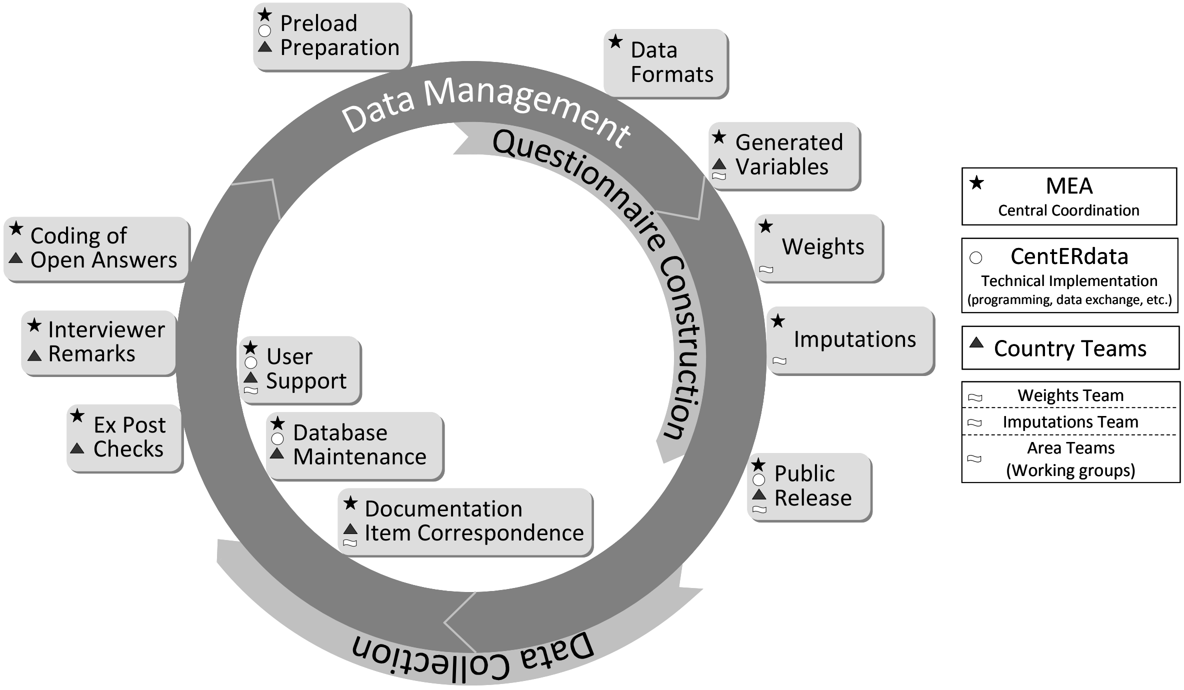
Data Management
Questionnaire Construction
Data Collection
★ ○ ▲ Preload Preparation
★ Data Formats
★ ▲ ≈ Generated Variables
★ ≈ Weights
★ ≈ Imputations
★ ○ ▲ ≈ Public Release
★ ▲ Coding of Open Answers
★ ▲ Interviewer Remarks
★ ▲ Ex Post Checks
★ ○ ▲ ≈ User Support
★ ○ ▲ Database Maintenance
★ ▲ ≈ Documentation Item Correspondence
★ MEA Central Coordination
○ CentERdata Technical Implementation (programming, data exchange, etc.)
▲ Country Teams
≈ Weights Team
≈ Imputations Team
≈ Area Teams (Working groups)

Figure 6.1: Database management tasks in SHARE and actors involved

6.2 Ex post checks

After the end of the fieldwork period the survey agencies from all participating countries send the CAPI data one final time to CentERdata who then upload these raw data to an internal server. One central task for the database management team at MEA, the country operators, and the survey agencies is to check and correct the delivered raw data and to prepare the data for the public release and for the preload of the upcoming wave. Thus, the central database management team has to run two parallel processes: first, producing a public release dataset for the current wave and second, generating a preload database for the next wave of data collection. For both processes it is essential to have the same basic checks and corrections on several variables, mainly on IDs, demographics, and household composition.

For a longitudinal study like SHARE, a correct matching of households and individuals across waves is essential: It is a precondition for analysing changes across time as well as for preloading necessary information into the CAPI instruments of following waves. On the individual level it is compulsory that correct IDs, demographic variables, information about moves and deceases, panel status (i.e. whether to use the baseline or longitudinal questionnaire), and other preload measures are linked to the right respondent. On the households level correct IDs are important for the matching with the survey agency address files. This section focuses on the matching of households and individuals across waves. To give the reader a notion of magnitude of the challenge: only in about 5% of the cases problems exist with the demographic information – however, these 5% constitute 95% or more of the work associated with the data cleaning process.

In a first step the database management team at MEA corrects IDs mostly on the household level according to indications by the survey agencies. Those get information about potential household mix-ups directly from their interviewers. Sometimes interviewers conduct an interview using a wrong household ID. This is either due to technical problems or interviewers simply click on the wrong line in the Sample Management System (SMS) when starting the interview. Survey agencies also have information about moves and deceases from their panel care activities which is used to correct the household composition for the next wave.

Secondly, MEA does systematic checks to identify mix-ups within households. These are detectable by merging the information gathered in the different modules of the CAPI-data of SHARELIFE (in the first two waves of SHARE also drop-off and vignettes questionnaires were used). The most frequent problem within a household is that partners are mixed up. This means that interviewers by mistake questioned household member A on household member B’s ID and vice versa. Those mixed up partners can be corrected using a Stata procedure that compares gender, year of birth, marital status and other relevant information. Another “within household problem” which arises sometimes is that respondents get a new ID because interviewers did not realise that the respondent already lived in this household in the previous wave. By checking the household composition across waves the same persons interviewed with different IDs can be detected. Further on, the data cleaning team uses information from the remarks, which comes directly from the interviewers. Some corrections in IDs and demographics is based on this information.

With SHARELIFE a more elaborated SMS was introduced, that already records changes in the household composition and basic demographics of household members in the first part of the interview, the “coverscreen wizard” (see also Chapter 3). The household composition and basic demographics are preloaded if the respondents agreed to this procedure. Interviewers then check with the household respondent if there are people missing, if a person actually never lived in the household, or if anyone has moved in between the waves. Interviewers can also correct year and month of birth, which helps to consolidate the household composition and the demographics. Introducing this comparison to the household composition of the last wave allows for easier linking across panel waves, such that these consolidated data are then the backbone for the public release and the preload data.

It should be mentioned that another source of information about non-matching respondents is to compare the data gathered in the CAPI interview with administrative data. For the German part of SHARELIFE, data was linked with administrative records of the German Pension Fund ("Deutsche Rentenversicherung", SHARE-RV). For the cases successfully linked demographic information according to this very reliable administrative database can be corrected. This is planned in other countries as well for future waves.

6.3 Dealing with remarks

During the interview, interviewers have the possibility to enter a remark. These remarks are associated with the question the respondent is answering at that point in time – even though the remark may not have to do anything with that question. As interviewer remarks prove to be an important source for data cleaning, interviewers were instructed to record additional information which could be useful to understand respondents' answers. For every question interviewers could add a remark in the CAPI software by pressing a special shortcut. The remarks are recorded similar to the text of an open question and are of course in the language the interviewer uses.

There are multiple kinds of remarks, each having different implications for the subsequent data cleaning process:

1. Typographic errors: These are the most important remarks as they include information on the answer being wrong due to an interviewer mistake. In this case the recorded answer can easily be corrected.
2. Explanations of given answers: If the respondent was not sure whether he or she understood the question correctly.
3. Additional answer categories: Some remarks point to missing answer categories, e.g., if respondents feel that none of the provided answer options applies to their situation. These remarks are very helpful for the questionnaire design of following waves.
4. Problems during the interview: These remarks can help to give the survey agencies and the SHARE team feedback about potential sources of problems.
5. Other remarks

Potential of the remarks

One can distinguish two different potentials of the remarks, which tie back into the scheme of providing release and preload data: correction of erroneous actual data and improvements for the following waves. Data correction is mainly based on those remarks describing typographic errors. If the interviewer mistyped an answer and was not able to correct it during the interview, the data can be corrected ex post based on such a remark. Remarks explaining given answers could also be relevant for data cleaning. If a remark clearly indicates that a given answer is wrong and the right answer is also included, the remark serves as a source for data correction. However, the SHARE policy also employed in SHARELIFE concerning corrections based on interviewer remarks is very conservative: data will be corrected only if one can be absolutely sure that the given answer is wrong and if the right answer can be inferred from the remarks.

Knowledge of the problems which arise during the interview is very important for improving instrument and interviewer training for the next wave. There may be several reasons for such problems: Many remarks for the same question in one country may hint to translation problems of this question. If the remarks arise in different countries they may indicate that the question is unclear in general. On the one hand, this information can be used to change some questions for the following waves. On the other hand, it highlights which aspects of the questionnaire are unclear for respondents and should be included in the interviewer training for the next wave.

Remarks in SHARELIFE as an international survey

The total number of remarks differs from country to country. They range from less than 300 (Austria) to more than 4.000 in Sweden. But the countries also differ in the number of realized interviews. Figure 6.2 shows the average number of remarks per interview per country in SHARELIFE.

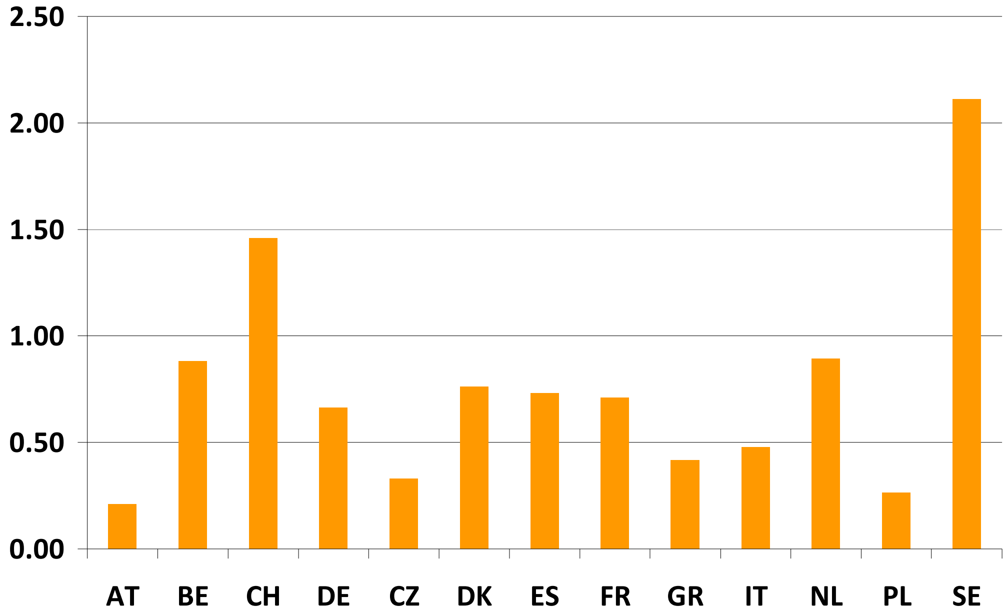

Figure 6.2: Remarks per interview across SHARELIFE countries

Except for Sweden and Switzerland, all countries have on average less than one remark per interview. The reasons for the country differences in the total number of remarks are unknown yet, but might point to more problems during the interview in some countries (e.g. due to translation issues) or to differences in interviewer training across countries (see also Chapter 5).

Due to the fact that SHARE is an international survey, the handling of remarks cannot be accomplished centrally for all countries as the remarks are in the respondents' respective national language. However, it is very important to harmonize the handling of remarks across all countries, because dealing differently with remarks may lead to unwanted variation in the data across countries. Therefore, the process requires cooperation between the central coordination and each country team. In order to achieve a standardized procedure the data cleaning team in Mannheim has created a programme to handle the interviewer remarks for all countries.

Country teams receive MS Excel files with all their country's remarks which include a "*Stata do-file generator*". (Most country teams and the central data cleaning team in Mannheim use the statistical software package Stata^{TM} for data management and cleaning. The MS Excel sheet is designed to produce standardized Stata code lines.) The main work is to decide, module by module, which remarks are relevant for the central data cleaning team, the country teams, or the survey agencies. If remarks allow for direct correction of the data, country team operators fill in the different columns in the file and the corresponding Stata command is automatically generated. In most of the cases, the remarks are useful to identify problematic questions, but the data cannot be corrected. For example, remarks often include explanations why none of the possible answer categories fit with the answer the respondent wanted to give. In this case the correct answer is available but one cannot correct the data because the needed answer category is not included. Here, the country teams add a "flag variable" with the (translated) respondent's answer, which is then collected and compared over all countries by the central data cleaning team in Mannheim.

6.4 Coding open answers

SHARELIFE data, as well as the previous SHARE waves, include a number of variables storing text information (string variables). Most often these variables contain respondents' open answers to follow-up questions. For example, questions with categorical answer options often include a category "other" that allows for specification. Another frequent type of question resulting in text information in SHARELIFE is the currency corresponding to an amount of money stated in a previous question. Variables of this kind have to be processed before they can be included in a public release. The raw data may contain sensitive information that allows for inferences on the persons interviewed. Thus, variables containing open answers have at least to be screened and cleaned in this respect. Moreover, such variables are not very user friendly, especially as the text information is stored in the many different languages used for the interviews in the participating countries. For this reason, one would like to code open answers into existing or new categories, which are in some cases used in a following wave as a regular answer category.

The most extensive coding in SHARELIFE was done for the currency information. Here, the problem is even more severe, because the raw currency strings are difficult to use right away. SHARELIFE allowed respondents to report financial amounts in the currency in which they can give the most accurate account. Interviewers then simply typed in the respective currency as a string. There are three reasons why the editing of these strings must be performed very accurately: First, the life-spans of SHARELIFE respondents in most countries cover various currencies, e.g. due to currency reforms or migration. It can, however, not be assumed that respondents report amounts in the currency of the respective time and country: they might also convert an amount corresponding to a pre-Euro time into Euros or vice versa. Second, respondents often supplied non-standard local currency names or abbreviations or they did not specify the country (e.g. “francs” may refer to France, Switzerland or Belgium). Third, currency abbreviations are often short (3 or less letters) which makes identifying typographic errors difficult.

For these reasons a routine was developed to code the raw currency information into a numeric variable using a code scheme of the most common current and former currency notations. The applied procedure started with creating country specific lists with unique strings extracted from *all* variables containing currency information. These lists were distributed to the country teams together with a code scheme of the most common current and former currency notations. Country teams then coded the strings into numeric variables as conservatively as possible using the supplied code scheme (i.e., no assumptions, not the “likeliest” currency). Only unambiguous answers were coded into the currency code scheme, additional codes were used for ambiguous strings.

A similar procedure was applied to the coding of open answer information as it appears in “other – specify” variables (e.g. question AC010, *other private residence*, or question AC012, *other non-private residence*). Again, lists with unique string information were produced and distributed to the country teams who assigned codes according to a provided coding scheme. These schemes contained the generic categories of the preceding question as well as additional categories that emerged after a screening of the data. The process of recoding open answers is ongoing – while some of the variables have been coded and included in a public release, further coding operations will be designed in collaboration with the country teams and interested data users to allow for tailored and ready-to-use generated variables.

6.5 Getting the Data Released

Getting a ready-to-use dataset released is what all the work on SHARELIFE and every other SHARE wave eventually results in. However, it is not an end point of the data cleaning process but rather a successful completion of a stage. User feedback on potential problems in the data or on the linkage of respondents across waves, for example, may bring up issues that call for further editing and a re-release of the data. This is one reason why the whole data processing in SHARE – from raw data to public release – is organised as a sequence of procedures controlled by one single master programme. Specifically, the master programme is a *Stata do-file* that calls a series of other do-files which include the actual data editing routines and commands. This setup guarantees that replication is always possible – in the case that something changes in the input data or along the way, one can always track the changes through the array of programmes.

Besides issues of data quality a user-friendly data structure is also of importance for a publicly available data set. User-friendliness involves an additional editing of the cleaned raw data. Therefore, not only data cleaning programmes but also formatting routines are addressed by the master programme. These routines include

- the creation of sets of dummy variables to store answers where a respondent can choose more than one of several answers (e.g. question AC018, question CS007);
- the conversion of any specified amount to Euros (waves 1 & 2) using current conversion rates for non-Euro countries and fixed conversion rates for pre-Euro information in Euro countries;
- the editing of "unfolding bracket" variables (see *SHARE Release Guides* on the SHARE website) holding financial information: auxiliary variables are eliminated and a consistent naming and labelling structure is applied;
- the assignment of consistent missing values and non-response codes (-1/-2 for "don't know"/"refusal", -9999991/-9999992 in case of financial variables that might take negative values) as well as variable label information.

In addition to the edited data from the CAPI interview, SHARE release data is supplemented by modules including generated variables, such as the body mass index, depression scales (Euro-D), or ISCED codes, as well as sampling design weights, calibrated cross sectional weights and calibrated longitudinal weights. Furthermore, multiply imputed values are available for a set of demographic variables, individual and household level economic variables, as well as generated variables.

6.6 Documentation and user support

When the scientific use data are released, one of the most important parts is the documentation and user support. Due to its longitudinal, cross-national and multidisciplinary nature, right from the start SHARE was a very large and complex research database requiring extensive documentation and user support. The provision of supplementary modules – as weights, imputations, and several topic-specific generated variables modules – further intensifies the complexity. With the release of SHARELIFE, focussing on people's life histories, another dimension of complexity is added. To assist researchers in efficiently using all parts of the database, the documentation concept was revised and is now organized in a three-part documentation structure. First, the revised "*SHARE Guide to Release X*" is designed as the core overview on all aspects of the released data, and is always held up-to-date. Second, to document country specifics, e.g. deviations from the generic questionnaire and their reasons, three interactive *"Item Correspondence"* tools complement the questionnaires, which are of course available in all country/language versions used. Finally, *"tailored user support"* is provided by both, the central and the country specific user-support teams. Besides these three main sources of information on the data, the SHARE homepage additionally contains an FAQ section, a newsletter and a publications archive. In the latter, complete versions of the previous and current *First Results Books* as well as the methodology volumes can be downloaded.

The *SHARE Guide to Release X* documents the relevant information for directly working with the released datasets. It covers basic information on participating countries, eligibility rules, the additional drop-off questionnaires and vignette studies, as well as general issues on the composition of data sets and types of respondents. More important for data analysis are the chapters on merging the data across different modules and of course across panel waves, and on how to merge SHARELIFE data and the two preceding two panel waves of the SHARE project. Furthermore, the Guide covers the treatment of missing codes, conversion of currencies into comparable Euro values, and the conversion of unfolding bracket questions and of multiple answer questions into "dummy" variable sets. Additionally, the SHARE Guide to Release provides information on specific issues, e.g. the coding of open answer/other questions in various modules, or on how to work with the selected child in the CH module, or the coding of nationality and country of birth. Finally, the multiple documentations on generated variable modules are now integrated into the guide (either as chapter in the main part or as appendix to the guide). One document now holds all information on the additionally generated datasets on weights, imputations, housing, health, social support & household composition, and alive-status. This includes also the documentation of the ISCED, ISCO, and NACE coding.

The most basic part of documentation of a survey project is the originally used *questionnaires*. SHARE provides for all waves and country/language versions the originally used instruments. They come complete with all technical details, i.e. filter rules, interviewer instructions, accepted answer ranges and looping rules. For an easier overview on country specific deviations from the generic questionnaires interactive *Item Correspondence* tools provide structured overviews of deviations within a wave and deviations in the generic version across the waves. These tools are available from the homepage only and can generate custom views on single countries, modules or questions. Currently, there are two cross-sectional correspondence tools available for the two released waves of SHARE (waves 1 & 2) that document country specific deviations; and a third tool for longitudinal changes in the generic questionnaire between these waves. Integrated in these tools are always English translations of all deviations. For single-country deviations from the generic version the respective country team also provides explanatory notes for the specific reasons.

A complementary release of an easy-to-read codebook is planned for the SHARELIFE data for the first time (a simpler version is also available in the appendix to this book). This will mark the final step in the revision of the documentation structure of the whole SHARE-project. Unlike the questionnaire, this codebook will be based on the released data rather than on the CAPI instrument. It then refers to variables as they are actually distributed and thus includes generated variables and documents all editing of the raw data. Apart from that it will preserve features of the questionnaire (question text, interviewer instructions, and, most important, routing information) but presented in a more clearly arranged way. The introduction of a codebook is a response to user requests as many of them found it hard to trace the elaborate filtering in the original questionnaires.

The third pillar of helping the scientific community to exploit the richness of the data is *tailored user support*. The central database management team in Mannheim as well as all country teams maintain email hotlines. Hence, user support is provided by those members of the SHARE team who implemented the surveys in each country, as well as by those who were in charge of producing the released data versions. The vast majority of user questions are directly answered within less than a week by the central Mannheim team as well as by the country teams. Questions on special issues or on generated datasets are directed to the appropriate team within the SHARE workgroups. SHARE also organizes scientific user conferences. Here, members of the various SHARE teams and researchers are present and provide assistance and comments.

6.7 Concluding Remarks

Any survey has to process data from its raw state coming directly from the field up to the point when they can be released to the scientific community. This process has become easier to some extent with the introduction of more advanced technology, for example the move from paper and pencil interviewing to computer assisted interviewing. In principle, the SHARE data collection effort is conducting the same survey in multiple countries, each of which with its own specific issues and challenges. Combining these multiple surveys into one large enterprise then results in more than just the sum of the parts – both in terms of the outcome for the scientific user and in terms of the data management work involved.