ZA5073

Kinship and Social Security (KASS)

- Technical Report -

Background: outline of the KASS project, and how the data in this archive relates to it

The purpose of this note is to

- describe the KASS project
- explain which data and research materials are being made available and under what conditions
- indicate further sources of information about the KASS project

Origin and form of project

Kinship is at the heart of European society, sharing with the state the primary responsibility for welfare and social reproduction. But the workings of kinship and their connections to state policy remain controversial. Politically the boundaries between state and family responsibilities are subject to constant challenge and renegotiation. Academically, received ideas have had to be revised in the light of social and demographic change, and of new theoretical developments and research in disciplines as various as sociology, demography, economics, anthropology and history.

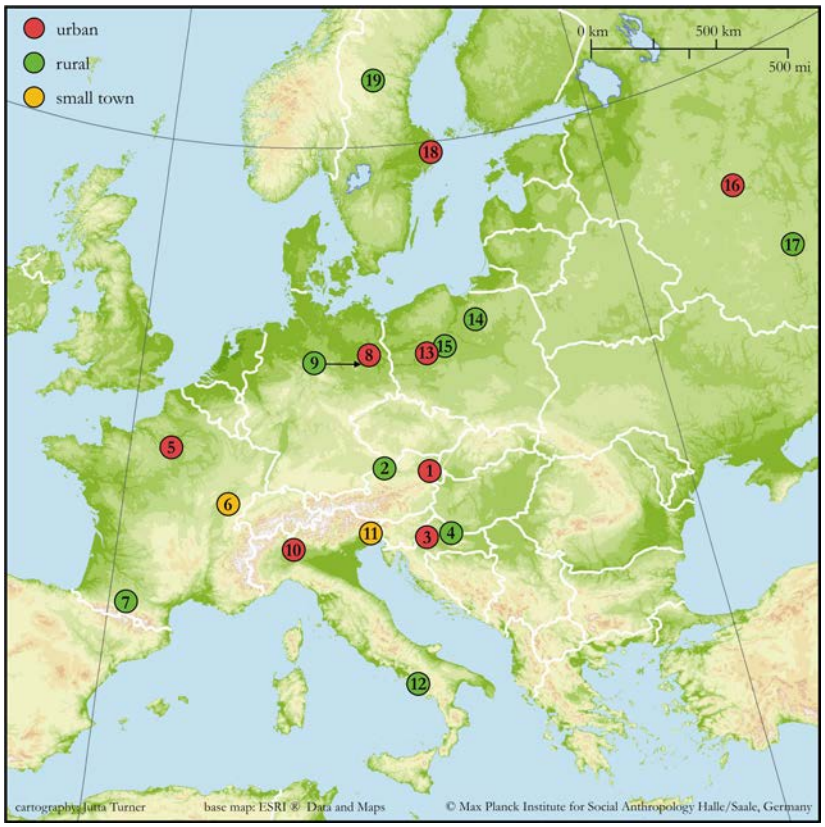
The data stored in this archive was collected as part of a multi-national research project on *kinship and social security* (acronym KASS), designed to contribute to these ongoing debates. The idea underlying the study was to look at the social security arrangements of families themselves – the extent to which people assist their closer or more distant relatives – in a set of different contexts which might help to explain the reasons for, and the consequences of, these practical arrangements. In doing so, the project drew particularly, but not exclusively, on two academic disciplines: history and social anthropology. Methodologically the project combined case-studies with comparative analysis.

The project itself was divided into three parts. The first part of the project focused on the history of the family during the twentieth century, in the context of political struggles over the welfare state, gender roles and parental authority. The case studies are at national level and concern eight European countries – Sweden, France, Germany, Austria, Italy, Croatia, Poland and Russia – with varying historical and cultural backgrounds as well as contrasting experiences during the twentieth century itself. Drawing on existing sources of historical, sociological and demographic data, the case study reports analyse the contribution of economic transformation and social policy to changes in family life – and debate whether these changes should be understood as a weakening or as a redefinition of traditional kinship roles.

The rest of the project consisted of two linked comparative studies, carried out together in nineteen research localities – two or three in each of the eight countries referred to above. The first of these studies used ethnographic methods to produce a qualitative picture of family relationships, the practical constraints under which they operate, and the support and control offered by wider kinship and community ties. The second study, for which the data was collected by the same research teams, was based on formal interviews with small but representative samples of informants from each locality. During the interviews a computerised questionnaire was used to collect quantitative data on the informants' kinship networks, and on the extent of mutual assistance and social interaction between each informant and other members of his or her network.

This archive contains the quantitative data collected using the computerised kinship network questionnaire – along with the questionnaire program and other programs which prepared the data for statistical analysis.

Map and table 1: The KASS field sites



No.	Country	Fieldsite name	population density per km²	% in agriculture full- and parttime (KNQ sample)
1	Austria	Karl Marx Hof (Vienna)	2,769	6
2	Austria	Schönau	49	30
3	Croatia	Travno (Novi Zagreb)	3,947	7
4	Croatia	Bjelovar	50	62
5	France	Nanterre (Paris)	6,914	0
6	France	Dole	650	47
7	France	Canton d'Aurignac	12	81
8	Germany	Marzahn (Berlin)	4,050	0
9	Germany	Glindow	279	4
10	Italy	Milano	6,998	0
11	Italy	Manzano	222	23
12	Italy	Tramonti	169	74
13	Poland	Poznań	2,168	0
14	Poland	Kurzętnik	58	70
15	Poland	Dziekanowice	48	41
16	Russia	Moscow	14,298	20
17	Russia	Kalikino	50	100
18	Sweden	Vällingby (Stockholm)	4,828	0
19	Sweden	Härjedalen	1	11

Contributions to KASS study and rights over data

KASS was funded by a grant from the European Union's Sixth Framework research programme. Additional financial support was given by the Max Planck Society (which coordinated the project) and by some of the other participating institutions.

The contracts underlying the project – between each institution and the European Union and between the members of the consortium itself – did not include any obligation to make the data publicly available. Nevertheless, it was agreed by most consortium members, and by the EU research directorate that it would be desirable to share the network data with the wider research community, subject to the restrictions that would be necessary to protect the privacy of the people who participated in the network interviews.

It was clearly also necessary to obtain the consent of each of the fieldwork teams who collected the network data. At the present time, the research teams for 15 out of the 19 field sites have agreed that their data can be made available for research purposes – and these are the sites whose data is included in this archive. (The sites concerned are those numbered 1-6, 10-13 and 15-19 in table 1 above.)

Contributions to KASS programming and rights over the KASS programs

The data in this archive was collected using the **Kinship Network Questionnaire** (KNQ). This was developed from a program called Kinship Editor, written by Professor Michael Fischer of the University of Kent in Canterbury, England – who kindly gave us access to his original Java code. The program for the KNQ itself was developed by Gordon Milligan and Christian Kieser in line with the specifications in the KASS research proposal and feedback from the research teams on a pilot version.

In order to analyse the kinship network data it had to be converted from the original XML data files to matrix format, and a large number of composite variables needed to be derived. This was done in two sets of programs: "Download and Variable Derivation" (DAVD), and a series of variable-derivation programs written in SPSS code. (More information is provided in ZA5073_computing-guide in Folder II).

The data-conversion part of DAVD was originally specified by Michael Schnegg and Tuba Bircan, and the variable derivation part was specified by Patrick Heady. Tuba Bircan integrated these into an overall system, with help from Christian Kieser, and some further work on the system was done by Rob White. The programs were restructured and extensively documented by Zhonghui Ou, to whom the current version is due.

The data-conversion parts of the DAVD programs incorporate version V1.08 of XmlParser, a free software program written by Frank Vanden Berghen in C++, which provides standard algorithms for converting XML files to text matrix data.

Nearly all of the SPSS programs were written and documented by Siegfried Gruber – drawing in some places on contributions by Michael Schnegg and Patrick Heady.

These programs (both in compiled form and as non-compiled program code) are available to any interested user free of charge – subject to the conditions that users

1) acknowledge the authorship stated here
2) and, if they develop further programs using the code provided here, make their own programs freely available on the same terms.

Sources of information about the KASS study

"Family, kinship and state"

The results of the study were published in three volumes in 2010

- 1 H. Grandits (ed.), *Family, kinship and state in contemporary Europe, Vol. 1: The century of welfare: eight countries*, Frankfurt and New York: Campus,
- 2 P. Heady and P. Schweitzer (eds.), *Family, kinship and state in contemporary Europe, Vol. 2: The view from below: nineteen localities*, Frankfurt and New York: Campus,
- 3 P. Heady and M. Kohli (eds.), *Family, kinship and state in contemporary Europe, Vol. 3: Perspectives on theory and policy*, Frankfurt and New York: Campus

Users of the data are advised to consult these volumes for the results of the study – and also for a fuller description of the theoretical background than can be provided here. In this archive we refer to these books as FKS Volumes I to III.

The historical part of the project is reported in Volume I.

The reporting of the two linked comparative studies is divided between Volumes II and III although, because of the close connection between them, cross-references are frequent. The first chapter in Volume II outlines the theoretical background and methodology of both studies, and the second chapter uses data from the network study (i.e. the data stored in this archive) to provide some initial quantitative comparisons between the nineteen different research sites. The rest of Volume II is devoted to ethnographic accounts of family and community life in each of the nineteen localities – rounded off by an ethnographically based discussion of the interconnections between kinship and community ties.

In the first part of Volume III the authors use statistical analyses of the network data to measure the extent of mutual assistance between relatives and explore its connection to residence and marriage patterns, intergenerational relations, gender roles and fertility. They go on to review the findings of the whole project – linking them with the findings of other research and drawing critically on theories of altruism, reciprocity, cultural continuity and socio-economic change. The concluding chapter reviews the preceding discussion and makes a number of policy recommendations. Methodological appendices provide reference points for some of the more technical aspects of the comparative network research.

"Ethnologie Française"

Half of the January 2012 issue of Ethnologie Française was devoted to the ethnographic and network aspects of KASS. The issue includes several ethnographic reports in Franch and English, as well as English-language accounts of the overall methodology and some of the main quantitative findings.

Additional information is available on the project websites.
www.eth.mpg.de/kass
www.eth.mpg.de/kass/eu

Patrick Heady
Max Planck Institute for Social Anthropology
June 2014

Introduction: how to use the files in this archive

The files in this archive are organised into four folders.

I. **ZA5073_folder_I** provides an overview of the whole project, including background information and methodology
II. **ZA5073_folder_II** contains the documentation needed to understand the data-files, including an overview of the computing procedures
III. **ZA5073_folder_III** contains programs and program documentation – including both the computerised Kinship Network Questionnaire (KNQ), and subsequent programs for transforming the data and deriving new variables
IV. **ZA5073_folder_IV** contains the data files themselves.

Folders I and II can be accessed directly over the web.

Folder III is available on request, subject to agreement to abide by the rules of creative commons when using or adapting the programs it contains.

Because of its detailed and confidential nature, the data in Folder IV can only be viewed under controlled conditions. There are two possibilities:

- access at the GESIS secure data centre at Köln,
- receipt of a copy of the dataset to be used at the analyst's own institution – subject to appropriate guarantees of confidentiality. (This option is only available to former members of the KASS research team.)

The rest of this note outlines the content of each folder.

Folder I: Overview of background and methododology

Including this introduction, there are seven documents in the Overview folder

- Introduction (ZA5073_introduction)
- Background (ZA5073_background)
- Methodology 1: research design and statistical issues (ZA5073_methods_1)
- Methodology 2: implementation and data quality (ZA5073_methods_2)
- Methodology 3: KASS fieldwork guide (ZA5073_methods_3)
- Methodology 4: KASS sampling guide (ZA5073_methods_4)
- Methodology 5: methods check-list (ZA5073_methods_5)

Background paper

The data in this archive was produced as part of the KASS (Kinship and Social Security) project. The background paper introduces the KASS project; explains how the data presented here relates to the project as a whole; outlines the conditions on which the data

and programs are available to users; and lists published sources of information on KASS methods and results.

Methodology papers 1 to 5
The first two methodology papers outline the essential features of the KASS network interview methodology – highlighting points that need to be taken into account during analysis. Many of these are issues that would arise in any survey – but there are also some important issues that are specific to KASS's network design.

Thus **ZA5073_methods_1**'s account of the research design covers the standard statistical issues of sampling, weighting, and the applicability of statistical tests. But it also discusses the various different populations – of persons, unions and relationships – to which the network data can refer, and the fact that there are often alternative ways of sub-sampling the data in order to estimate the characteristics of the same underlying population. It also discusses the implications of the fact that the primary sampling units were anthropological field-sites, which were selected purposively instead of by a random procedure.

The list of contents is as follows

- Theoretical aims of the study
- Universe (a.k.a. target populations)
    1. geographic scope
    2. units of analysis
    3. network relationships
- Data collection strategy
- Sample design
    1. selecting localities
    2. choosing respondents
- Statistical implications of the research design
    1. random sampling and exchangeability
    2. weighting procedures for informants and households
    3. sampling and weighting for other units and relationships
    4. spatial scales and hierarchical data levels
    5. modelling and testing
    6. geographical representativeness

Similarly **ZA5073_methods_2**'s account of implementation and data quality includes non-response levels, internal cross-checks for biased reporting, and comparisons of certain key variables with figures from other sources. However, it also points out that the implications of reporting "errors" may depend on the purpose of the analysis. Thus the clear evidence that relatives who are more than three genealogical steps away from our respondents are under-reported is a limitation the data from a demographic point of view; but it is also a valuable indication of the limits of kin-awareness from a socio-cultural point of view. In a similar way, close examination of discrepancies between levels of help given and received reveal not just sampling and reporting biases, but also conceptual issues concerning who people regard as the key beneficiaries of different kinds of collective help.

The list of contents is as follows

- Implementation
  1. interviewing procedures
  2. local sampling strategies
  3. non-response
- Representativeness of the achieved sample
  1. individual characteristics
  2. locality characteristics
  3. national and macro-regional characteristics
- Data quality
  1. network size
  2. domestic help
  3. other variables: inheritance and income

For further background on the way the data was collected, users can also consult

- **ZA5073_methods_3**: the guidance notes on fieldwork methods issued to the KASS field teams
- **ZA5074_methods_4**: a the sampling guidance notes issued to the field teams on ways of selecting the informants for kinship network interviews

ZA5073_methods_5 provides a short (one page and a bit) check list of the methodological issues that researchers should take into account when planning their analyses. Each point is followed by a reference to the places in ZA5073_methods_1 and ZA5073_methods_2 where the issue concerned is discussed, or to the computing guide and questionnaire text in Folder II.

Folder II: Data documentation

This folder contains

- Computing and dataset guide. (ZA5073_computing_guide – see outline below)
- Two text versions of the questionnaire
  - ZA5073_questionnaire_text_EN (in English)
  - ZA5073_questionnaire_text_DE (in German)

  Besides listing the questions themselves, the questionnaire texts include continuity prompts and indicate which members of the kinship network each question applies to.
- The SPSS data-file ZA5073_variables.sav, which sets out codes and labels for each variable in the main data file – but without the actual data. In effect this SPSS file provides a codebook, ordered in the same way as the main data-file itself.

Outline of Computing and Dataset Guide

The purpose of this guide is to

- describe the main features of the SPSS data files, including both
  - the main individual-level data-file
  - the localities data file
- explain how the datasets were constructed, describing in turn
  - the KNQ interview, and how to view the XML files which it produced
  - the conversion of the XML data to matrix format, and the creation of key network variables
  - the SPSS variable derivation programs
  - data editing
  - the creation of the localities dataset
- enable users to generate comparable datasets from their own field research
  - by using the existing KASS programs
  - by modifying the programs to fit their own research designs.

The guide goes through the issues in this order, focusing on the main points. Where more detail is required, readers are referred the supporting documentation in Folders II and III.

Folder III: Programs and program documentation

Three kinds of program were needed to create the KASS database. First was the KNQ, the kinship network questionnaire program which the interviewers used to collect data from our informants. Next came the DAVD ("download and variable derivation") programs, which rewrote the KNQ data into a format that could be used for statistical analysis, and also derived some basic analytic variables from the network structure. Finally came a sequence of programs written in SPSS programming language which derived many more variables from the downloaded data, producing the data sets which were used for the analysis reported in the second and third volumes of "Family, kinship and state" (see ZA5073_background). Each set of programs is stored – along with relevant documentation – in its own sub-folder:

- ZA5073_folder_IIIa_KNQ
- ZA5073_folder_IIIb_DAVD
- ZA5073_folder_IIIc_SPSS.

KNQ programs and documentation

Folder IIIa is divided into three parts

- **KNQ_1 compiled programs** contains the computerised questionnaires that were actually used for interviewing in the languages of the eight KASS countries:
  knq_AT_1.03
  knq_DE_1.03
  knq_FR_1.03
  knq_HR_1.03
  knq_IT_1.03
  knq_PL_1.03

knq_RU_1.03
knq_SE_1.03
along with an English-language version of the same program
knq_EN_1.03

- **KNQ_2 non-compiled master program** contains the master-program which was used to generate each of the national questionnaire programs. It is in open-code format and can be modified to add translations into other languages, or – with more difficulty – to change the set of underlying questions and procedures.

- **KNQ_3 documentation** contains three files
  - ZA5073_KNQ_description_1 is a general introduction to the kinship network questionnaire
  - ZA5073_KNQ_description_2 is the "KNQ user guide" which explains to interviewers how to use the compiled KNQ programs.
  - ZA5073_KNQ_description_3 is the "KNQ Functional Design Specification" which explains the structure of the non-compiled master program. Along with the Java documentation which forms part of the program code, it provides the information needed to adapt the KNQ program for use in other studies.

DAVD programs and documentation
Folder IIIb is also divided into three parts.

- **DAVD_1 compiled programs** contains the sequence of DAVD programs in the form of executable (".exe") files with GUI interfaces. These files can be run directly to transform the files produced by the KNQ into matrix format suitable for reading by SPSS, and to generate some network-related derived variables.

- **DAVD_2 non-compiled programs** contains the same programs in the form of C++ code. If the KNQ is adapted for use in other studies it would also be necessary to adapt these programs to allow for changes in the initial KNQ-generated data-files – **unless** the change was simply a matter of language, in which case no change would be needed.

- **DAVD_3 documentation** contains two elements.
  - ZA5073_DAVD_description_1 gives an overview of the DAVD programs.
  - ZA5073_DAVD_description_2 is a folder which contains PDF files of the main elements of the C++ programs – setting out the code itself alongside explanatory comments.

SPSS programs and documentation
Folder IIIc is divided into two parts.

- **SPSS_1 programs** contains a suite of 33 programs (numbered 0 to 32) which further develop the output of the DAVD programs. The SPSS programs can be run in sequence to generate the variables that were used in the SPSS based analyses

reported in “Family kinship and state” volumes 2 and 3 (see reference in ZA5073_background). The programs are written in SPSS programming code and are given in non-compiled form – which means that the SPSS package is needed to run them.

- **SPSS_2 documentation** contains the file
  - ZA5073_SPSS_description_1 which gives a connected overview of the suite of SPSS variable derivation programs.

Folder_IV: Data files

This folder contains the data sets themselves.

- **ZA5073_folder_IVa** contains the anonymised XML data files for each interview

- **ZA5073_folder_IVb** contains the main SPSS data set

- **ZA5073_folder_IVc** contains the SPSS data set for localities analysis.

Patrick Heady
Max Planck Institute for Social Anthropology
June 2014

Computing and Dataset Guide

Most users of the KASS data held by GESIS will start by using the SPSS data files. This guide has three aims.

1) To help users get started, the first part of this guide describes the main features of the two SPSS data-files – and gives some guidance about how to use them.

2) For users who want to explore the data in greater depth, the guide goes on to explain how the data was collected and processed in order to produce the SPSS data files. As part of this, it introduces
   - the documented program of the Kinship Network Questionnaire (KNQ)
   - the original data files in the visual XML format in which they were collected
   - the programs that converted this data to matrix format and derived some basic variables describing the kinship network
   - the SPSS programs that created most of the derived variables
   - the ways in which data values were checked and edited.

3) The final part of the guide is addressed to people who would like to use KASS methods to generate datasets in their own field research. It explains how to do this
   - by using the existing KASS programs
   - by modifying the programs to fit their own research designs.

The guide itself will present the main points. Where more detail is needed, it will refer to the other documents in which the information is provided.

In order to use the data effectively, analysts should also consult the methodology guidance notes in **Folder I** – particularly **ZA5073_methods_1** and **ZA5073_methods_2**, and the check-list in **ZA5073_methods_5** – for information on statistical and data quality issues.

Part 1: Introduction to the SPSS datasets

The data for the study was collected by interviews with selected individuals, using the computerised **Kinship Network Questionnaire** (a.k.a. **KNQ**).

The KNQ collects data about
- the informant's network of known kin,
- the "unions" to which the network members belonged (i.e. families of origin and partnership/procreation, see **ZA5073_methods_1** for definition)
- "important others" (see definition in **ZA5073_methods_1**),
- mutual assistance at present and over the informant's life-time,
- information that may help us to understand the patterns of mutual assistance, including
- background information about resources and needs, and about help received from the state and official organizations

- information about social interactions and cognition (i.e. the words people use to classify their relatives).

Text versions of the KNQ questionnaires can be found in **Folder II**. The programs themselves, along with explanatory documentation, are stored in **Folder IIIa**.

The data collected in these interviews forms the basis of the two SPSS data files contained in **Folder IV**:

- The individual level file, **EnPersonUnionData-18c_2010_15_localities.sav**, contained in **sub-folder IVb**.
- The locality-level file, **aggr1fieldsite_2010_15_localities.sav**, contained in **sub-folder IVc**

1(a) Main features of the individual level SPSS file

The complete individual level SPSS data set used for the analyses in the FKS (the series of KASS-based books) is called **EnPersonUnionData-18c_2010_15_localities.sav** . It contains 2691 variables, including both the original variables collected in the interview itself, and a large number of derived variables – produced by combining the original variables in various ways.

It also contains data on 40,220 individuals – 444 informants ("egos") living in 15 field sites, along with 39,776 other members of their networks. [The full 19-field-site data-set used for the analyses reported in Volumes 2 and 3 of "Family. Kinship and state in contemporary Europe" contained data on 50,676 individuals – 570 informants and 50,106 other members of their networks.]

In this section we describe

I. the thematic structure of the 2691 variables (data columns)
II. the hierarchical structure of the sample of individuals (data rows), and how to choose the right sub-samples and weighting systems for particular analyses
III. the naming system for the derived variables on domestic help, and how to choose the correct combination of variable, sub-sample and weighting system

1(a)(i) Thematic structure – finding the right variables

We have followed SPSS conventions in giving each variable an explanatory "variable label" and, in the case of categorical variables, "value labels" as well. The codes and labels for each variable are set out – but without the actual data – in the SPSS data-file **ZA5073_variables.sav** in Folder II. In effect this SPSS file is provides a codebook, ordered in the same way as the main data-file itself

The variables are divided into several thematic groups – which are listed in Table 1 below, along with the column numbers in which the variables of each group appear.

Table 1: Groups of variables and their position within the data files

	Order	no. of variables
general and identifier variables	1-13	13
original variables	14-299	286
re-entered variables	300-324	25
re-classified variables	325-350	26
derived network matrix variables	351-372	22
basic additional variables	373-405	33
ego's data	406-414	9
relationship variables	415-571	157
weighting variables	572-581	10
union variables	582-852	271
contact variables	853-879	27
wedding etc. variables	880-915	36
dwelling help and building help	916-986	71
major gifts	987-1016	30
inheritance	1017-1168	152
residence variables	1169-1207	39
illness help	1208-1283	76
household structure	1284-1329	46
income, contributions to shared household finances, renting the home	1330-1386	57
regular financial contributions from outside the household and major loans	1387-1468	82
farming help	1469-1590	122
personal problems, emotional closeness	1591-1640	50
childcare help	1641-1758	118
domestic help (inside household)	1759-1977	219
domestic help (outside household)	1978-2279	302
domestic help (inside and outside of household)	2280-2691	412

The section headings should be fairly self-explanatory, but the following remarks may help.

- the **"identifier variables**" in the first section can be used for data-inspection, to select subsamples for analysis, and for various operations in the process of deriving new variables. The following comments relate to each of these aspects.
  - **KNQid** contains the filename of the XML file, and **pid** is the person identifier generated during the interview itself. The **pid** number also appears by each person's icon when the KNQ program is used to open the XML files included in this archive. For this reason these two variables are very useful when comparing the data in the SPSS file with the original XML data. However, they are less suitable for scrolling through the SPSS database itelf, since KNQid is hard to read, and they are gaps in the pid sequence, due to mistakes and deletions during the interview itself.
  - The SPSS data file is organised in country order, then by field-site within country, by interview within field-site, and by individual within interview. The corresponding identifiers at each level are **countryId**, **fieldsite**, **KNQsid2**, and **apid**.
  - When selecting subsamples for analysis, one often needs to restrict the sample to informants ("Egos") or else to exclude Egos. **apid** can be used for this purpose, since Ego always takes the value 1. Alternatively the variable **if_ego** can be used. Another variable which is often needed for subsample selection is **in_egos_household**, though this occurs later in the data-file (in column 389).
  - **person_id** (based on the values of KNQsid2 and apid) identifies each individual uniquely in the whole dataset. This is useful for some repetitive aspects of variable derivation.
  - for more information see **ZA5073_DAVD_description_1** in Folder IIIb and **ZA5073_SPSS_description_1** in Folder IIIc.

- **"original variables"** are the answers recorded in the KNQ to the questions about individual characteristics and interactions with Ego.
  - The questions are listed in the English and German text versions of the KNQ questionnaire (in **Folder II**), which also indicate which members of the kinship network each question applies to.
  - Note that the order in which the variables are stored in the SPSS data file is different from the order in the questionnaire. In particular the data from KNQ section 8 is stored before the data from sections 4 through 7.

- **"re-entered variables"** are mostly quantitative variables which were originally recorded in free format. The re-entered versions are in a standard format. (See subsection 2(d) below).

- **"reclassified variables"** are versions of variables which were originally defined for SPSS purposes as "string" variables, but which have been reclassified as "numeric". Later sections of the database also include variables which have been reclassified in this way. (See subsection 2(d) below).

- **"derived network matrix variables"** are the variables derived during the conversion process from XML to matrix data. These include

- the kinship path that linked Ego to each of the individuals to whom s/he was genealogically connected. This path was stored in two ways – (a) as a sequence of personal identifier codes specifying the individuals who formed the path linking Ego to the individual question and (b) as a sequence of relationship codes specifying the connection (father, mother, brother, sister, son, daughter, husband, wife) between each successive pair of individuals along this kinship path
- the number of different kinship paths linking Ego to the individual concerned (if there was more than one path, the information about the connecting links referred to the shortest path)
- a set of "demographic" variables specifying the number of children, the number of parents (i.e. 0, 1 or 2), the number of siblings, and the number of partners reported for each individual. [Note that these figures are the joint product of the demographic facts themselves and the informants' awareness and willingness to report them. The analytical implications of this are discussed in **ZA5073_methods_2** pages 7 to 11)
- some overall measures referring to the interview as a whole: the total number of individuals recorded; the total number of recorded kin (i.e. people connected to Ego by descent or marriage), and the total number of "important others".

- **"ego's data"** include nine items of information about Ego which have been entered as separate items in the rows for each individual in the interview concerned. This makes it easier to cross-tabulate the characteristics of *Egos* and *Alters*, and also to derive new variables which depend on the characteristics of both.

- **"weighting variables"**: as **ZA5073_methods_1** explains, it is important to weight KASS data correctly. This section contains some variables used in calculating the weights, and concludes with the two weighting variables themselves: **"ego_weight"** and **"hhld_weight"** corresponding respectively to "weighting system 1" and "weighting system 2" described in **ZA5073_methods_1**.

- **"union variables"** includes the identifier codes for each union to which the individual concerned belongs, as well as a number of variables giving information about the union in question. The dataset allows for each individual belonging to anything up to six unions.
  - **union1_id** to **union6_id** contain the identifier code for each of these six unions, which consists of the **usid** identifier generated by DAVD for the union concerned, preceded by the value of the interview identifier KNQsid2. These codes identify each union uniquely within the data set as a whole, though of course the same value will occur in the data rows for each of the other individuals belonging to the union concerned.
  - For more information about the derivation of these codes see **ZA5073_DAVD_description_1** in Folder IIIb and **ZA5073_SPSS_description_1** in Folder IIIc.

- **"domestic help"** is a collective title for eight kinds of help which can be given to, or received from, people inside or outside ego's household. In order to cover the different possibilities in a comparable way, we developed a standard naming system, which is explained in section 1(a)(iii) below.

1(a)(ii) Hierarchical structure: choosing the right sub-sample and weighting system for each analysis

The way the data can be used depends on the research design, which is described in **ZA5073_methods_1**. Users should consult the section on "statistical implications of the research design" – particularly the discussions of "weighting procedures for informants and households", "sampling and weighting procedures for other units and relationships" and "spatial scales: theoretical questions, spatial hierarchies and hierarchical data levels".

The strategy chosen depends on two factors, the "data level" of the variables concerned, and the population which you wish to represent.

- "**Interview level**" data consists of variables which have one value for the whole interview. Examples would be Ego's characteristics (since there is only one informant/Ego for each interview), the characteristics of Ego's household (e.g. the number of people it contains), or the characteristics of Ego's network (e.g. the number of relative he named).
- "**Network member level**" data consists of variables that can take a different value for every member of Ego's network (including Ego herself). Examples would be either personal characteristics such as age-group or occupation; information about the kinship tie between the person concerned and Ego (brother, daughter's husband, or whatever); or information about the individual's interactions with Ego (whether they have been in contact in the last month, whether either has helped the other with farm work in the last year, and so on).

It is important to keep in mind how the two kinds of data are stored in the database.

- Network member level data is listed in the row for the individual concerned. (Note the information about relationships and interactions with Ego are not stored in Ego's row, but only in the row for the other network member).
- Interview level data is usually held on all the rows for the interview concerned (both on Ego's own row, and the row for every other network member. (An alternative procedure would have been to hold interview-level data only on Ego's row; but that would have prevented us from relating network member variables to data about Ego's characteristics, or about the network as a whole.)

The fact that both kinds of variable occur on all rows of the data matrix, means that special attention must be given to choosing both the right variables, and the right sub-sample for the analysis concerned. To obtain a representative result, it is also important to weight the data in the right way.

Table 2 in **ZA5073_methods_1** sums up the recommendations about how to combine interview level data, network member level data, and the two different weighting systems to represent either…

A. adults (using informants (Egos) as a sample)
B. households (using Egos' households as a sample)
C. relationships (using the relationships between Egos and other individuals as a sample)
D. individual people of all ages (using household members as a sample) .

The relevant information for analyses A and B is at "interview level" – which means that only the data rows for Egos should be included in the analysis. As noted above, these rows can be selected using the variables **if_ego** or **apid**. For A the data should be weighted using **ego_weight**. For B, **hhld_weight** should be used.

For C the subsample should include every row except the rows for Egos – since Ego does not have helping relationships with him- or herself. The subsample can be obtained by using the variables **if_ego** or **apid** to exclude the data rows for Egos (i.e. using them in the opposite way to that for analyses A or B). **ego_weight** should be used to weight the data.

For D the analysis applies to the subsample of individuals (including Ego herself) who live in each Ego's household. The data rows concerned can be selected using **in_egos_household**. The weighting variable to use is **hhld_weight**.

1(a)(iii) The naming system for the "domestic help" variables

"Domestic help" is a collective title for eight kinds of help which can be given to, or received from, people inside or outside ego's household. In order to cover the different possibilities in a comparable way, we developed some standard naming conventions. Table 2 illustrates the conventions for variables relating a specific kind of help (taking cooking as an example). Table 3 sets out the main features of the naming system for variables dealing with domestic help in general (i.e. help with any of the eight domestic tasks.)

Table 2 Variable names connected with a specific domestic task (cooking)

	kind of help	**in household**	**outside household**	**Overall**	**variable refers to…**
		from KNQ section 4	**from KNQ section 5**	**from KNQ sections 4 and 5**	
1	ego gives help in a specific way (to a particular individual) yes no	cooking_ihh_e2o	cooking_e2o	not calculated	ego-other relationship
2	ego receives help in a specific way (from a particular individual) yes no	cooking_ihh_o2e	cooking_o2e	not calculated	ego-other relationship
3	ego gives help in a specific way (to anyone) yes no	cooking_ihh_e2o_yes	cooking_e2o_yes	cooking_e2o_all_yes	ego
4	ego receives help in a specific way (from anyone) yes no	cooking_ihh_o2e_yes	cooking_o2e_yes	cooking_o2e_all_yes	ego
5	Person helps with household task yes no	cooking			ego and other household members
6	Contribution of men and women to household task	who_cooking			ego's household as a whole

Table 3 Variable names for domestic help in general (calculated by combining the full set of domestic help variables)

	kind of help	in household	outside household	overall	variable refers to…
		from KNQ section 4	from KNQ section 5	from KNQ sections 4 and 5	
1	Ego gives practical help (to a particular individual) yes no	practical_ihh_e2o	practical_ohh_e2o	practical_e2o	ego-other relationship
2	Ego receives practical help (from a particular individual) yes no	practical_ihh_o2e	practical_ohh_o2e	practical_o2e	ego-other relationship
3	Ego gives practical help (to anyone) yes no	any_practical_ihh_e2o	any_practical_ohh_e2o	any_practical_e2o	ego
4	Ego receives practical help (from anyone) yes no	any_practical_ihh_o2e	any_practical_ohh_o2e	any_practical_o2e	ego

A glance at the first four lines of each table shows that the two sets of conventions are similar but not the same. In both cases the naming system allows for three main distinctions.

1. Between help given by Ego to someone else (marked "e2o"), and help received by Ego from someone else (marked "o2e").
2. Between variables that record interactions between Ego and specific individuals, and variables which aggregate this information, showing whether Ego exchanges this kind of help with anyone in the network. The names of the first kind of variables, which we can call "relationship" variables (see final column of table), are not marked in any particular way. But the ego-aggregated variables are marked – by the suffix "yes" in the case of the task-specific variables in Table 2, and by the prefix "any" in the case of the general-help variables in Table 3.
3. Between (a) help which Ego exchanges with people in the same household, (b) help which ego exchanges with people in other households, and (c) variables which include both kinds of help. (a) is always indicated by the suffix "ihh" – but the treatment of (b) and (c) differs between task-specific and general-help variables. In the case of task-specific variables (b) is not indicated in any particular way, and (c) is either not calculated, or else indicated by the suffix "all". In the case of general-help variables (b) is indicated by the suffix "ohh", while (c) is not indicated by any particular suffix.

The use of two different indicators – "yes" and "any" – to mark ego-aggregated variables has no particular significance; it simply reflects the fact that the task-specific and general-help variables were programmed at different times. However the fact that – in connection with distinction 3, the unmarked version of the variable name refers to exchanges of help

with people outside Ego's household in the case of task-specific variables, but all exchanges of help in the case of general-help variables – reflects a real change in our analysis strategy.

The changes relates to the treatment of data on help exchanged within Ego's household. A glance at the text questionnaires (in **ZA5073_questionnaire_texts** in **Folder II**) or the compiled versions (in **Folder IIIa**, sub-folder **KNQ_1**) shows that different questioning procedures were used for exchanges of help within Egos household, and exchanges of help with people outside Ego's household. In the latter case each individual who provided help, or received help from Ego, is explicitly recorded. However in the case of help within Ego's household we merely asked which members of the household (including Ego) provided help with each of the tasks concerned, but did not ask specifically ask about the recipients of the help.

Because of this our initial analysis strategy was to restrict the analysis of inter-personal exchanges of help to exchanges between Ego and people in other households. This is reflected in the absence of any particular suffix to indicate exchanges of help with people outside the household – since this was the only kind of help being considered at that point – and is the reason why this convention was used for the task-specific variables, which were programmed before the general help variables. Later on, we made some assumptions about which individuals benefited from the help provided within the household, which made it possible to include within-household help in the overall analysis of help exchanged between Ego and other network members. The "ihh" and "overall" variables in the first four rows of both tables are based on these assumptions. In the case of the general-help variables we changed the naming system to reflect the new approach. But we did not change the names of task-specific outside-household variables which had already been programmed.

However there are also some questions about domestic help for which the original form of within-household helping data is more helpful. This is indicated in the fifth and sixth rows of Table 2. The untransformed variable "cooking" (in row 5) provides information on the contribution of each member of Ego's household. Data of this kind can be aggregated to household level to provide information about the division of domestic labour in each household – an example of which is the variable "who_cooking" which records whether cooking for other household members is done by women only, by men only, or by both sexes.

As we saw in the previous section, 1(a)(ii), specific subsampling and weighting strategies are needed to achieve representative results different kinds of variable. The final column of each table notes the kind of unit each variable refers to, and thus which sub-sampling and weighting strategy applies. The appropriate strategies are

A. for variables labelled "ego" in the final column
B. for the variable labelled "ego's household as a whole"
C. for variables labelled "ego-other relationships"
D. for the variable labelled "ego and other household members".

The names of other variables in the domestic-help sections of the database extend the principles described here in various ways – for instance, to allow for the intensity of each helping relationship, or to summarise Ego's exchanges with different categories of relative.

However, the principles set out here, combined with the more detailed information provided in the variable and value labels, should enable you to interpret the variables concerned.

One final important point is that the way one analyses the data on domestic help depends on the assumptions one makes about the beneficiaries of each reported act of assistance. One aspect of this has already been mentioned – namely the problem of identifying the beneficiaries of help provided by people within Ego's household. Despite the different question format, problems also arise in connection with help provided to, and received from, people outside Ego's household. Both sets of problems, and the assumptions we made, are discussed in **ZA_5073_methods_2**, in the section on "help with domestic work".

1(b) Localities (field-site) dataset

The other dataset – **aggr1fieldsite_2010_15_localities.sav** – is a file in which the cases are the localities (field sites) themselves, and the variables provide information about those localities. These include

I. 151 variables derived (by aggregating and averaging) from the data in the individual-level data-file (**EnPersonUnionData-18c_2010_15_localities.sav**)
II. 12 variables derived from other statistical sources
III. 7 locality identification and classification variables

The full 19-locality version of this dataset was used to construct the locality-level statistical charts in the second and third volumes of *Family Kinship and State in Contemporary Europe* (referred to below as FKS).

1(b)(i) Aggregated variables derived from the individual level SPSS file

The variables concerned are listed in rows 2 to 152 of the data file. All of them provide summary information based on the interviews carried out in the field site concerned. The variables are of three kinds.

a) Summary data about the sample expressed as totals.
b) Mean (average) values based on sample data.
c) Proportions based on sample data.

Since (b) and (c) are intended to be representative of the field-site populations (subject to the limits of sampling error), the data on which they were based had to be selected and weighted appropriately – following the principles set out in section 1(a)(ii) above, and explained in **ZA5073_methods_1**. Details are contained in the program description and text.

1(b)(ii) Variables derived from other statistical sources

Rows 153 to 158 of the data file contain figures derived from local official statistics. In each case we chose official statistics for the area most closely corresponding to the field-site concerned – though the match between the official area and our *de facto* field site was

never exact. Details of the local official sources are given in Appendix 4 of FKS Volume 3 (pp 432-437).

Rows 163-166 contain national statistics for the countries in which our field sites were located. The sources used for this data are also indicated in the appendix just mentioned – and fuller information is provided in FKS Volume 1's Appendix on data sources and derivations (pp391-400).

Rows 167 and 168 contain national-level data on attitudes to friendship, obtained from the European Values Study (1999) and the International Social Survey Program (2001). For further information see FKS Volume 1's Appendix on data sources and derivations (page 391).

1(b)(iii) Identity and classification variables

The seven variables concerned are located in rows 1, 159 to 162, 169 and 170.

- "locality_name" (row 159) names the local field site
- "region_agestructure" (row 160) names the area to which the local official statistics refer
- "fieldsite" (row 1), "fieldsite_a"(row 161) and "country_alt_org" (row 162) can be used in graphs to label the relevant data points.
- "fieldId_mean" (row 170) is an area-type variable, classifying each locality as "rural", "small urban" or "urban".

The final variable "kinship_terminology" (row 169) could also be called "Macro-Region". It classifies the eight countries included in the KASS study into three unequal groups.

1. Sweden
2. France, Germany, Austria
3. Italy, Croatia, Poland, Russia.

The classification is explained in FKS Volume 3 in two distinct but connected ways.

- In chapter 2 (pp 42-43) it is linked to patterns emerging from present-day statistics and from historical studies.
- In chapter 15 (pp351-355) it is linked to the structure of kinship terms in the languages of the countries concerned.

Part 2 How the data were produced

The production of the data sets took place in 5 stages

1. collection using the kinship network questionnaire (KNQ)
2. conversion of data to matrix format, and creation of key network variables
3. derivation of new variables
4. editing
5. creation of the localities dataset.

2(a) The KNQ interview

Text versions of the questionnaires (**ZA5073_questionnaire_text_EN** in English, and **ZA5073_questionnaire_text_DE** in German) can be found in **Folder II**. Besides listing the questions themselves, the questionnaire texts include continuity prompts and indicate which members of the kinship network each question applies to.

Folder IIIa contains the KNQ programs themselves, as well as some explanatory documentation.

- Compiled forms of the KNQ programs in English and each of the KASS languages are stored in sub-folder **KNQ_1.**
- The non-compiled master program can be found in sub-folder **KNQ_2**.
- Sub-folder **KNQ_3** contains three descriptions of the program:
  - **ZA5073_KNQ_description_1** gives a brief description of the underlying logic
  - **ZA5073_KNQ_description_2** is the "KNQ User Guide"
  - **ZA5073_KNQ_description_3** sets out the "KNQ Functional and Design Specification.

The KNQ data is stored in a format known as "XML". The files for the 570 KASS interviews are held in anonymised form (with personal names replaced by identification numbers corresponding to the variable **"pid"** in the SPSS datafile) in **Folder IVa**. The XML files can be viewed in two was.

- The XML code itself can be seen by using a text reading program such as Notepad or Wordpad to open the files.
- Alternatively the KNQ program can be used to open the XML data files, viewing them as they would have appeared during the interview itself.

2(b) conversion of data to matrix format, and creation of key network variables

"XML" is suitable for handling data collected over a screen, but is not suitable for statistical analysis. For statistical analysis the data has to be converted into matrix format. The programs which did this are known collectively as DAVD ("download and variable derivation"), and are described in **ZA5073_DAVD_description_1** in Folder IIIb.

Creation of key network variables

The DAVD programs also used the network data collected by the KNQ to calculate some additional variables which would be important for the later analysis. These were

- a set of identifier codes (described in part 1 of this guide) which make it possible to identify separately all of the 570 interviews, every individual referred to during each interview, and every "union" to which these individuals belonged
- a set of "matrix variables" (also described in part 1 of this guide)

Storing the data in matrix format

The data was set out in a two large matrices "EnPersonData" and "UnionPerson" containing the data from all the interviews. Each row of these matrices held the data for a single individual referred to during an interview. Data for the individuals belonging to the same interview are stored in successive rows, starting with the row for Ego (the informant).

EnPersonData held detailed information for each individual. It also contained summary information about the interview as a whole.

Each row of UnionPerson held information about the "unions" to which that individual belonged, specifically the identifier code for each union and whether the individual belonged to it as a partner/parent or as an offspring.

Both matrices were initially created in text format, and then converted to SPSS data files for further work. They were combined into a single matrix by the SPSS variable-derivation programs.

2(c) The SPSS variable-derivation programs

Most of the analysis reported in "Family, Kinship and State" volumes 2 and 3 (see **ZA5073_background**) uses derived variables which combined and condense the raw data collected during the KNQ interview. With the exception of the variables derived at the data-conversion stage, these new variables were all created by programs written in SPSS programming language.

A set of 31 SPSS programs was used to derive the additional variables included in the main SPSS dataset, and a 32nd program calculated the average values for the localities-level dataset. **ZA5073_SPSS_description_1** (in Folder IIIc) outlines the structure of this set of programs (including an initial "program 00" which brings the total to 33). It starts by discussing some issues that apply to the whole set of programs, and goes on to give a short summary of each – explaining which variables are derived in which part of each program.

The programs themselves (including some internal documentation) are also given in Folder IIIc. They should be of interest to users who wish to know exactly how the variables were derived – and can of course be amended to produce modified versions of the variables.

2(d) Data editing

We made a number of adjustments to the data in the SPSS files. The reasons for this were…

I. to reformat data that were not yet in a suitable form for analysis;
II. to correct implausible or impossible values
III. to check for the completeness of data on some topics and exclude cases with incomplete data from the relevant analyses
IV. to create new variables with imputed values based on the answers to existing variables supplemented by appropriate assumptions

After each editing process, new variables were created to hold the checked and revised data. No changes were made to the values of the original variables collected during the interview itself (and stored in columns 14-299 of the dataset). As a result the impact of the editing process can be assessed by comparing the original and revised values. If the edited results are felt to be misleading, it would be possible to go back to the original data and start again.

The main editing processes were as follows.

- Formatting quantitative data. During the interviews quantitative data (on dates, amounts of money, and proportions) was entered in free format. This free format data (which is shown in the columns for the original quantitative variables) had to put into a consistent format. This was done by hand, and the results fed into the data base (in columns 300-324) by SPSS program 02. During this process a few obviously incorrect values were changed.
- Changing the SPSS data-type classifications. For its own purposes SPSS requires each variable to be classified as “string” or “numeric” – and most items were originally classified by the DAVD programs as “string”. However, it is easier to program and analyse data which classified as “numeric”. For this reason, many of the SPSS programs start by creating “numeric” versions of the original variables which they are going to use.
- Most checks for incorrect or incomplete data were run in the SPSS programs dealing with particular subject areas. For instance, programs 18 “inheritance” and 22 “income” include important checks of these kinds.
- Two notable areas in which data were imputed were
    - Age in years, in cases where the year of birth was not available. If an age-group value was available for the individual concerned, Program 03 used the central value of the relevant age group to estimate the individual’s age.
    - Variables concerning the exchange of help between Ego and other members of her household (see section 1(a)(iii) above) were created in programs 26, 29 and 30.

2(e) Creation of localities dataset

The final SPSS program calculates average sample values at locality level for a number of variables, and stores these average values as variables in a file in which the cases are the localities themselves. We have also added a number of variables drawn from other sources: local census values, and average national values for some variables drawn from national statistics and from comparative social surveys.

The data set itself is described in section 1(b) above. Details of its derivation are given in the relevant part of **ZA5073_SPSS_description_1** and in the text of Program 32.

Part 3: Using the KASS programs to create new data sets

3(a) Using the programs in their existing form

The existing programs can be used to conduct interviews and generate new SPSS datasets. Four steps are involved.

1. As a first step, the compiled KNQs (stored along with user instructions in Folder_IIIa) can be used to create new XML data sets.
2. Next the compiled version of the conversion software (stored in Folder IIIb) can be used to generate text matrix data sets, and SPSS program files for converting these data sets to SPSS files.
3. The SPSS programs generated by the main conversion software can then be run to convert the new text matrix files into SPSS data files
4. Finally the suite of SPSS variable derivation programs in Folder IIIc can be run to create the equivalent of **EnPersonUnionData-18c.sav**.

For each stage of the process it is necessary to use a computer with the correct software environment.

- The KNQ requires a Java runtime enivironment.
- In order for DAVD to run, the 2008 version of Microsoft Visual Studio needs to be installed on the computer.
- In order to run the SPSS programs, the SPSS package needs to be installed on the computer.

3(b) Modifying the KASS programs

But researchers may not wish to use the programs in exactly their present form. They may wish

- to translate the questions into another language,
- or to modify the questions themselves.

To help future researchers do this we have also attached

- detailed program descriptions for the KNQ and the conversion programs (see **ZA5073_KNQ_description_3** in Folder IIIa, and **ZA5073_DAVD_description_1** and **2** in Folder IIIb)

- the program texts themselves in open non-compiled form (in the same folders)

Both of the documents just mentioned include sections which explain how the program can be adapted to include new or changed questions. Even so, amending either the KNQ or the DAVD programs would be a job for an experienced programmer.

It would also be necessary to have the right software environment in which to read and amend the programs, which were written in Java (KNQ) and C++ (DAVD).

Patrick Heady
Max Planck Institute for Social Anthropology
June 2014