German Council for Social and Economic Data (RatSWD)

RatSWD

Working Paper Series

Working Paper No. 84

Access to and Documentation of Publicly Financed Survey Data

Wolfgang Jagodzinski, Christof Wolf

May 2009

Federal Ministry of Education and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Access to and Documentation of Publicly Financed Survey Data[1]

Wolfgang Jagodzinski and Christof Wolf

GESIS-Leibniz-Institute for the Social Sciences (Christof.Wolf[at]gesis.org)

Abstract

The topic of this paper is access to and documentation of survey data financed through public funds. We distinguish between four types of publicly financed survey data: (1) Academic survey data from the national or international research infrastructures; (2) data from DFG projects or similarly funded projects; (3) survey data collected in research projects funded by the Federal State and the Länder (Ressortforschung); (4) Population and Household surveys from national and international statistical agencies. For each of these types of data we describe the current situation and present recommendations for future developments.

Keywords: Survey data, data access, data documentation, data archive

1 Anmerkung des RatSWD: Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder haben seit 2004 ein umfangreiches Datenangebot für die Wissenschaft aufgebaut, das über 70 Statistiken umfasst. Inhaltlich sind alle Bereiche vertreten, für die in der amtlichen Statistik Daten erhoben werden. Besonders intensiv werden die Bevölkerungs- und Haushaltserhebungen, die Wirtschaftsstatistiken sowie die Steuerdaten genutzt. Insgesamt wurden seit 2004 für rund 900 Forschungsprojekte Daten bei den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder beantragt. Um die Nutzung der Daten zu erleichtern, wird neben den Scientific-Use-Files zur Off-Site-Nutzung die so genannte On-Site-Nutzung angeboten. Hier erfolgt die Arbeit mit den Daten entweder an Gastwissenschaftlerarbeitsplätzen in den Statistischen Ämtern oder über die kontrollierte Datenfernverarbeitung. Die Wahl der Zugangswege hängt von dem konkreten Forschungsvorhaben und von der Anonymisierbarkeit der Daten ab. Die Haushaltserhebungen wie der Mikrozensus werden in der Regel über die Off-Site-Nutzung bereit gestellt. Für komplexe Forschungsvorhaben im Bereich der Wirtschafts- und Gesundheitswissenschaft ist hingegen meist die On-Site-Nutzung der präferierte Nutzungsweg.

1. Introduction: Four Data Types

Our recommendations refer to four data types: (1) Academic survey data from the national (like ALLBUS or SOEP) or international (like ESS, SHARE, ISSP, EVS, or CSES) research infrastructure; (2) data from DFG projects or similarly funded projects; (3) data collected in research projects funded by the Federal State and the States (Ressortforschung); (4) Population and Household surveys from national and international statistical agencies. We will briefly describe the current situation and make suggestions for future developments for each of these data types. We do not attempt to give a comprehensive overview over all existing survey programs, however. Special programs are discussed in the reviews of specific domains. We also do not address problems concerning register data.

2. National and International Research Infrastructure

2.1 Present Situation

Surveys under academic conduct which are part of the research infrastructure (national and international survey programs) are the main source of comparative studies either in a longitudinal or in a comparative perspective. In Germany national programs like ALLBUS and SOEP are seen as part of the research infrastructure for the social sciences and thus they are fully funded. With regard to international surveys the situation is more heterogeneous. As far as ISSP is concerned, the costs for the German survey as well as a large share of the costs for processing of the international dataset are seen as investments into the international research infrastructure and publicly funded. The European Values Study has recently reached a similar status. The last wave (EVS 2008) has been publicly funded and costs of data processing are divided between Tilburg and the GESIS data archive.

Panel studies like SOEP are optimally suited for analyzing change over time. They are not only expensive, however, but also require a highly developed infrastructure for data collection and data processing. It is therefore difficult to organize multi-wave panel studies on an international level. Apart from very few exceptions like SHARE the large international survey programs therefore still are cross-sectional. Most of them in the mean-time have built up sequences of cross-sections which permit cohort studies for the analysis of change. Standards for international surveys have recently been published by the ISR in Michigan (http://ccsg.isr.umich.edu).

There is high demand for these studies. It becomes visible in the large number of data downloads and distributed copies as well as in the publications. Almost all survey programs publish their own bibliography.

The demand also justifies larger investments in data documentation and data improvements. Some progress has been made in the standardization and harmonization of data. The European Social Survey has set new standards for the documentation of international studies. Several programs have started to add context data to the microdata files.

The continuous growth and improvements of the data base as well as the high demand of the scientific community both guarantee the application of the most recent technology of data processing and therefore an almost optimal access to the data. Even though some of these programs are based on a mixed funding they largely follow the recommendation of the OECD for fully publicly funded research data (http://www.oecd.org/dataoecd/9/61/38500813.pdf). In a few survey programs the time point of general data access is still a point of discussion. As long as primary investigators are also responsible for the national funding they sometimes postpone the open data access in time. However the situation has already considerably improved during the last years. This problem would immediately be solved on a contractual basis if an international infrastructure for academic survey programs could be established. ESS and SHARE to our knowledge so far are the only science driven survey programs which receive the funding of the overhead costs from an international organisation.

The other restrictions come from data protection laws. Datasets which are offered for free download on the internet therefore usually do not include fine-graded regional or occupational variables. A reduced version of the ALLBUS (ALLBUScompact) is freely accessible. Larger versions of the ALLBUS and of international social surveys like ESS, EVS or ISSP can be downloaded for free for scientific use. If data protection requires a special contract between the researcher and the user, data are distributed individually. The scientist has only to pay handling charges for data delivery.

2.2 Recommendations

It would be highly desirable if the data quality of other international survey programs could reach the quality of the ESS in the future. This would require, however, larger budgets for the international research infrastructure. The ESS has also set new standards for the documentation of sampling and data collection which should be gradually adopted by other programs. Furthermore, the translation process as well as its documentation can be improved. Until recently the translation of international surveys was under the responsibility of the

national teams and largely terra incognita for secondary analysts. They could only get the final questionnaire which often did not even include interviewer instructions. Recent developments attempt to reach a higher degree of standardization and transparency.[2]

Other activities would require the institutionalization of a larger international infrastructure which would not only advise researchers in data collection and data processing but also coordinate different survey programs. In particular the input standardization of socio-demographic variables should be achieved. It would also be desirable to improve comparability by including sub-modules of items from time to time into different research programs or by integrating different surveys into a common data base.

3. DFG Projects and other Scientific Projects

3.1 Present Situation

While the data access to publicly funded national and international survey programs which belong to the research infrastructure is fairly satisfying, the access to data of singular scientific projects funded by the German Research Foundation (DFG) and other comparable foundations still leaves quite a lot to be desired.[3] GESIS has recently attempted to identify DFG projects from the years 2003 to 2005 which probably meet the acquisition criteria of the GESIS data archive.[4] Due to the limitations of the project documentations it cannot be decided in all instances whether the project meets the criteria or not. What can be safely said however is that more than half of the studies which almost certainly meet the criteria are not sent to the data archive.[5]

Basic rules of scientific conduct require that data have to be made accessible for replications. However, they do not require delivering the data to an archive. On the one hand regarding the costs of archival work it is debatable whether all project data should be deposited in an archive. On the other hand there are serious doubts whether empirical data even if they have been stored on floppy disks or tapes years ago, still are accessible. The serious limitations of meta-analyses clearly show that access to the original data is always preferable over confining oneself to published results of statistical analysis.

---

2 Thus EVS 2008 has recently used the web-based translation module WEBTRANS developed by Gallup Europe for reaching a centralized control of the translation process, better comparability of the translations in different languages, more uniformity of the final questionnaires and better documentation for comparative analyses.

3 For a detailed description of the perspective of the German Research Foundation on the development of social science infrastructure see Nießen and Kämper in this volume.

4 In principle, the GESIS data archive only accepts representative studies of populations or larger subpopulations which are relevant to social science research. It does not acquire experimental studies, for instance.

5 The results can be obtained from the authors.

3.2 Recommendations

In our view modern information technology allow a substantial improvement of the present situation in two directions.

First of all, we propose to define a minimum standard of data accessibility which has to be guaranteed by all publicly funded scientific projects: All data have to be stored in a digital repository which is provided by the social science infrastructure. The researcher does not store the data on a disk in the university but in a domain which is kept by a publicly funded institution. The obvious advantage of this solution to the researchers is that they do not have to care about backups and data transfer onto new PCs. All these tasks are in the responsibility of the institution hosting the data repository. Special agreements between data producers and the hosting institution will address all questions concerning data ownership, data access and data distribution. The data producer is free to choose between different options, i.e. the rights to the data do not automatically go to the data host. The advantages offered by such a system should be an incentive for storing the data at a central place.

Second we should distinguish at least between two different types of project data, those which are only relevant to a small group of scientists and data of broader interest. For the former type of data a mode of *self-archiving* should be established. This is based on clear division of labour: The data are stored at a central place like the data archive in Cologne but data processing and documentation is done by the primary investigator. The social science infrastructure should provide the researchers with attractive self-storage tools which help them to document and preserve the data. These tools may allow lower and higher standards of data processing they may also enable the researcher to build up simple and more sophisticated data bases and to combine data and publications. However, the project has the main responsibility for data deposition and the data archive should not be involved to a larger extent in this process.

Clearly, a number of questions have to be clarified before a mode of self-archiving can be established. What exactly is the division of labour between the social science infrastructure and the primary investigators? Who is responsible for the migration of data to new computer systems? Who protects the primary investigator against the violation of laws, in particular laws of data confidentiality? What kind of facilitating tools for data processing should be developed?

Self-archiving and self-documentation are not sufficient for datasets, which probably will be of interest for a larger group of researchers. These data should not only be stored in the data archive, but they should be processed in accordance with the most advanced standards of

data processing and documentation. It is advisable to consult the archive already in the early stage in the project as it is done in all important international survey programs. The involvement of an archive requires additional resources. And these resources should be included in the cost calculation of the research project from the very beginning.

A near at hand objection to our proposal is that the distinction between data of restricted and broader interest is artificial and vague. For example: Hasn't it sometimes turned out that a study like the election study of 1953[6] which was almost forgotten in the fifties became extremely important for the analysis of long-term change in later decades? Yes, it happens from time to time. We think, however, that reviewers of project applications have a fairly good judgement whether a dataset will have the potential for secondary analyses or not. Collaborative research units, for instance, will usually produce datasets which are highly salient for the scientific community at large. And if half a million or more Euros are granted for a representative national sample it is often at least implicitly assumed that these data will not be used exclusively by the primary investigators. Details of the procedure have to be further elaborated too.

We therefore suggest a pilot project which further clarifies the terms and modalities of assisted self-archiving within a central data repository and professional data archiving. Such a project also should come up with proposals for self-archiving tools.

4. Research projects funded by the Federal Government or State Governments (Ressortforschung)

4.1 Present Situation

Research in this field is mainly carried out by Governmental Research Agencies (GRA) and partly by external researchers. The GRAs have recently been evaluated by a governmental research committee (see http://www.wissenschaftsrat.de/engl_rechts.htm#EVAL) of the German Research Council (Wissenschaftsrat). Besides the evaluation reports on 28 institutes the committee has published comprehensive "Recommendations on the Role and Future Development of Governmental Research Agencies with R&D Activities" in May 2006, January 2007, May 2008, and November 2008.[7] Further single reports and additional comprehensive recommendations are announced for 2009. As far as the service of research and development infrastructure (R&D infrastructure) and data access is concerned the Recommendations of 1 April 2007 on page 11 state:

6 ZA-Study number S0145, so called Reigrotzki-Study.
7 http://www.wissenschaftsrat.de/texte/7854-07.pdf.

- "All Federal Ministries and their agencies should avoid installing redundant and expensive R&D infrastructure. The R&D infrastructure should instead be subject to use by scientists from all kinds of R&D establishments. Such joint use requires that information on the infrastructure be readily available. Therefore, within the next two years, the BMBF in cooperation with all other federal ministries should compile a compendium listing all R&D infrastructure in GRAs (especially instruments and data). This compendium should be made available to all universities and research establishments in Germany. The Government is also advised to release scientific use files to research data centres, thus granting external scientists access to specific data collections. If such data centres cannot be created, other instruments such as work places for visiting scientists should be used to facilitate access." (http://www.wissenschaftsrat.de/texte/7854-07.pdf).

The establishment of research data centres at a subset of the GRAs will improve the accessibility of data to smaller or larger extents. Institutes like the German Youth Institute (DJI) have already delivered their data to the GESIS data archive in the past so that the scientific community will mainly benefit from the new working places at the institute and the access to single and cumulative data files which so far have not been made accessible. In other instances research data centres will lead to considerable improvements.

The committee of the Wissenschaftsrat so far has focused on the research of GRAs but quite a few of its recommendations either concern or also apply to research projects which are carried out by external researchers. We therefore need not to go into detail here but can confine ourselves to two minor issues which to our knowledge have not been systematically addressed.

The first is the scientific use file (SUF). Its production is expensive and requires technical and methodological skills which often are not available at a GRA. It is more difficult to provide the scientific community continuously with SUFs than to establish one or two work places for visiting scientists. As a consequence, SUFs might actually obtain a low priority in the emerging research data centres. Work places for scientists are not substitutes for SUFs, however, because the latter allow a more flexible and less time-consuming data analysis. They therefore act as a much lower barrier against secondary analysis than work places in remote institutions. The committee report neither lists potential SUFs, nor defines selection criteria, nor discusses the cost-effective production of SUFs. It is particular ambiguous in the latter respect: While the second last sentence in the upper quotation can be read in such a way that externally produced SUFs should be released to the new research data centres, the German

version by contrast defines the production of SUFs as a task of the research data centres.[8]

The second problem concerns the release of data from projects which are funded by the Federal or State Governments. While some government departments, in particular the *Bundesministerium für Familie, Senioren, Frauen und Jugend*[9] follow a fairly open policy, others are more restrictive. There is no general regulation so far.[10] If research projects of this type become visible in the media the GESIS data archive directly approaches the primary investigators. Sometimes these attempts are successful and the data are acquired by the archive. Quite a few datasets, however, never become accessible for the scientific community.

4.2 Recommendations

Our recommendations focus on the two topics previously mentioned. As far as SUFs are concerned we share the preference of the German Science Council. In order to secure an optimal number of SUFs, experts should first ascertain the demand for SUFs and define priorities. If the SUF-priority is sufficiently high, the most cost efficient mode of file production has to be determined. SUFs can be produced by the research data centre alone, or in close co-operation with an experienced external organization, or by an external organization alone. It can be distributed by the research data centre, by the external organization or by both. Looking at the recommendation of the German Science Council and its English translation from this perspective they point to two different modes of SUF-production: While the German text aims at the SUF-production by a research data centre at the GRA, the English translation alludes to the SUF-production by an external agency. Both interpretations are correct insofar as the cost efficient solution will differ from GRA to GRA. There presumably is no general solution to the problem. In any case it is highly desirable that the cost efficient production of SUFs in this area is tackled as fast as possible.

The question of data release should be investigated more systematically by the committee of the Wissenschaftsrat. In our view, the previous considerations should hold: If data from *Ressortforschung* are in the interest of the scientific community they should in general be accessible. Rules of data confidentiality which are often seen as an obstacle to data access actually are rarely a reason for withholding a complete dataset. More often they only require

---

8 „Im Rahmen von Forschungsdatenzentren sollen ‚scientific use files' erstellt werden, die externen Wissenschaftlern die Auswertung ausgewählter Datensammlungen erleichtern sollen. Wo ‚scientific use files' nicht möglich sind, sollen die Forschungsdatenzentren mit Hilfe anderer Instrumente (z.B. Fernrechnen und Gastwissenschaftlerarbeitsplätze) Daten auf geeignete Weise zugänglich machen."

9 Negotiations between the Zentralarchiv (now: GESIS data archive) and the Bundesministerium für Familie, Senioren, Frauen und Jugend have resulted in the decision that data of research projects which are funded by this government department are regularly delivered to the GESIS data archive at the end of the project. The datasets which the archive obtains are usually of high quality and well documented.

10 The Eurobarometers are another example of publicly funded surveys which are regularly delivered to the GESIS data archive.

the cut off of some information and variables. In addition, access to sensible data may be offered in safe data centres. Free access to data for scientific purposes, in any case, should be the general rule and exceptions should be allowed only in a few, well-founded instances.

5. Household Surveys from Official Statistics

Large scale data collections produced under the auspices national statistical agencies have specific strengths that make them especially interesting for social and economic research. With respect to population or household surveys the large sample sizes and the usually very low non-response rates make these data a valuable source for economic and social-structural investigation.[11] They are regularly used for purposes of social monitoring, e.g. in the Datenreport (Statistisches Bundesamt et al. 2008), or for the construction of social indicators, e.g. Education at a Glance (OECD 2007) or the European System of Social Indicators EUSI.[12] However, these data are also used for a wide range of different analytical purposes. See, e.g., the extensive bibliographies of articles based on the Scientific Use Files of the German Labour Force Survey or the German Income and Expenditure Survey.

5.1 Present situation

The most important household surveys for socioeconomic research from official statistics in Germany are the Mikrozensus, the Einkommens- und Verbrauchsstichprobe and the Zeitbudgeterhebung.

The Mikrozensus – Germany's Labour Force Survey – is an annual random sample survey of one percent of Germany's population. It has been carried out in West-Germany since 1957 and in reunified Germany since 1991. Integrated into the Mikrozensus is the German part of the European Labour Force Survey. Because participation in the Mikrozensus is obligatory response rates are close to 100 percent. With over 800.000 individuals it is the largest population survey in Europe.

Germany's income and expenditure survey is called "Einkommens- und Verbrauchsstichprobe". This survey is conducted every fifth year since 1963. The survey is based on a quota sample, participation is voluntary.

The Zeitbudgeterhebung is Germany's time use survey. It was conducted for the first time in 1991/92 and repeated 10 years later in 2001/2002. The Zeitbudgeterhebung is a quota

11 Other data from official statistics include business surveys and process produced data; these are dealt with in other chapters in this volume.

12 http://www.gesis.org/en/services/data/social-indicators/eusi/

sample of over 12,000 individuals living in 5,400 households. The questionnaire of the survey complies with Eurostat's recommendations for time use surveys, participation in the survey is voluntary.

In addition to these databases microdata from the censuses of 1970 and 1987 (West-Germany) and from 1981 (East-Germany) are currently available or will shortly be available for academic research.

Generally there are four different ways to access German microdata from official statistics:

- For most of the surveys mentioned above Scientific Use Files (SUFs) can be ordered from the Statistical Office by academic or research related institutions for the purpose of predefined scientific research purposes. Usage within these institutions is not restricted to German nationals though each individual working with a SUF has to be registered as data user by the Statistical Office. SUFs are microdata files that have been reasonably anonymized, i.e. anonymized in such a way that any identification of individuals is only possible by excessive expenditures of time, costs, and personnel (Federal Statistic Act 1987; Wirth 2008). This is typically achieved by providing only a subsample of the original dataset, e.g. 70 percent as in the case of the Mikrozensus, deleting most of the regional information and collapsing categories with small frequencies (see also Müller et al. 1995).[13] For the Einkommens- und Verbrauchsstichprobe there are currently data from seven years, the first from 1962/1963, the latest from 2003. For the Mikrozensus a total of 21 SUFs are currently available, the earliest coming from 1973, the latest from 2006.[14] The data from the two waves of the Zeitbudgeterhebung are also available as SUFs.
- A second possibility to access data from official statistics is provided by the research data centres of the Statistical Offices, the so called On-Site use.
- Thirdly, there is also the possibility of remotely accessing the official microdata. In this case the analyst provides syntax to the research data centres of the Statistical Offices, the research data centres executes the syntax and checks if the output complies with data confidentiality requirements. This form of access is especially valuable if direct access to microdata cannot be granted due to problems of data

13 Alternatively, when detailed regional information is kept, other attributes such as occupation, industry or nationality are recoded into larger categories (see Wirth et al. 2005).

14 The SUFs are created by the statistical agencies in close cooperation with the German Microdata Lab of GESIS in Mannheim (see Lüttinger et al. 2004; Schneider and Wolf 2008).

confidentiality. However this kind of problem refers mainly to establishment data but does not usually pose a problem for household or population data. If, however, the researcher wanting to work with data does not have the possibility to obtain a SUF e.g. because he or she is not working at a national research organization then remote access might be a helpful service.

- Finally, the statistical agencies provide so called Campus Files which are Public Use Files (PUFs). These files are absolutely anonymized and can therefore be used without restriction. They are especially useful for training purposes. With respect to household surveys there are currently three Campus Files for different waves of the Mikrozensus available from the website of the Research Data Centre of the Statistical Offices.[15]

According to a recent survey among users of German microdata from official statistics scientists clearly prefer SUF as mode of data access. *All* respondents have used SUFs. In addition a fifth has made use of remotely processing the data and 10 % have accessed the data in one of the research data centres of the statistical agencies (Lüttinger et al. 2007).

More and more researchers are interested in international comparative research. Regarding this growing demand, official microdata provided by Eurostat – the Statistical Office of the European Communities – comes in the focus. Eurostat currently provides access to microdata of four household surveys. These are the European Community Household Panel (ECHP), the European Union Labour Force Survey (EU-LFS), and the European Union Statistics on Income and Living Conditions (EU-SILC) (for a broader overview of European data see Elias in this volume).

The ECHP is a panel survey that started in the 12 member countries of the European Union in 1994 and continued on an annual basis until 2001 (8 waves; some additional countries joined the survey after its initial launch). The survey covers a wide range of topics concerning living conditions including detailed income information, financial situation in a wider sense, working life, housing situation, social relations, health and biographical information of the interviewed. The ECHP was Eurostat's attempt to create a comparative database following the principal of input harmonization (for the different harmonization strategies see below and Ehling 2003; Granda and Wolf 2009).

The European Union Labour Force Survey is a rotating random sample survey covering the population in private households in currently 30 European countries. The sampling units

15 http://www.forschungsdatenzentrum.de/campus-file.asp

are dwellings, household or individuals depending on the country-specific sampling frames. The collection of microdata, i.e. individual data, started in 1983. Since 1998, the EU LFS has developed into a continuous quarterly survey. The EU LFS is conducted by the national statistical institutes across Europe and is centrally processed by Eurostat. However, it follows an ex-ante output harmonization approach. The main aim of the EU-LFS is to provide comparable information on employed, unemployed and inactive persons of working age (15 years and above) in European countries. The definitions of employment and unemployment used in the EU-LFS closely follow the International Labour Organisation's guidelines.

EU-SILC is an annual statistic and was launched in 2004 in 13 Member States. From 2005 onwards the data are available for all EU25 Member States plus Island and Norway. Romania, Bulgaria, Turkey and Switzerland have launched EU-SILC in 2006. EU-SILC provides cross-sectional and longitudinal microdata on income, poverty, social exclusion, living conditions and health. It can be viewed as a successor of the ECHP, though it employs an ex-ante output harmonization approach. The reference population of EU-SILC is defined as all private households and all persons aged 16 and over within private household residing in the territory of the Member States at the time of data collection.

Other datasets initiated by the European Union or coordinated by Eurostat are either not available as an integrated microdata file or they are not distributed by Eurostat even though these data are of great interest for social research (for details see the next section).

5.2 Recommendations

Among the manifold challenges we face with respect to further developments in the field of population and household surveys from official statistics three seem to be especially pertinent from the perspective of socioeconomic research: further improvement of data access, adjustment of procedures to anonymize new data sources, enhancement of inter-temporal and cross-national comparability of data.

The improvement of data access can be divided into an improvement of documentation to ease access to data already available to the research community and the generation of access to new data sources. As is true for all secondary research, analyses of official microdata also depend on extensive documentation of the data and the data generation process. In addition, to be useful this information has to be formatted in a standardized form and organized in such a way that it can be accessed seamlessly (a document that is stored under a pile of other documents and that can be only read with a pair of "magic glasses" obviously is of no use). An example for a very thoroughly documented statistic is the German Mikrozensus. The

microdata information system MISSY developed by GESIS combines all available metadata for this survey and offers them in a coherently organized form through a web based system (http://www.gesis.org/MISSY; Janßen and Bohr 2006).

Data access should also be improved with respect to information on field procedures. Compared to what we know about the process of data collection in social surveys, e.g. the European Social Survey, field work procedures applied by the different Statistical Offices of the German Länder or in the different national offices in the EU are mostly terra incognita. For example, for the German Mikrozensus the number of interviewers, their workload, the number of call backs, non-response incidents and reasons for non-response are all unknown. However, at least for the Labour Force Survey the situation has improved over the last ten years. Today we know the mode of interviewing (self-administered, CAPI or CATI), the date of the interview and if the interview is a proxy interview.

A big problem still is access to data sources collected under regulations of or at least coordinated by the European Union. Currently only microdata from the above mentioned EU-LFS, ECHP and EU-SILC are available for research outside of Eurostat. Other data such as the Adult Education Survey, the Time Use Survey, Household Budget Survey, Statistics on Information and Communications Technologies (Household survey part) or Europe's Health Survey are currently not available for comparative research. If the Lisbon goal of the European Council should be met, namely Europe becoming by 2010 the "most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion", then research monitoring this progress is mandatory and this research needs access to the relevant data.

A new challenge for data access is posed by register data that will become more important in the next years. Here problems of integrating data from different registers and from registers and surveys has to be solved (Alda et al. 2005). Furthermore the currently applied methods of data anonymization have to be adapted to these new data sources. However, this is not totally new terrain.

A last issue of necessary improvements of microdata bases from official statistics that we like to address here is that of inter-temporal and especially cross-national comparability. At present EU data is collected on the basis of regulations detailing the variables that member states have to provide to Eurostat. This approach, called ex-ante output harmonization (Ehling 2003), leaves the concrete process of data collection to the data producer, i.e. each country has its own questionnaire and applies their own field procedures. This flexibility of data collection makes it easier for the national statistical offices to integrate the data collection process into

their national programs and particularities. The comparability of data for demographic and socioeconomic variables yielded by this approach is generally satisfactory. This is especially the case where international standard classifications such as ISCO or NACE are available and the countries agree on their interpretation and application. However, even with such "factual" information as highest educational degree (Schneider 2008) or supervisory status (Pollak et al. 2009) output harmonization may lead to incomparable data. Naturally this is much more true for subjective data such as health status, life satisfaction or happiness, all of which are included in the EU-SILC program.

The analytical potential of microdata collected under EU regulations and integrated by Eurostat could be improved without larger costs if the following three recommendations were applied: Firstly, although it might not be feasible and for some variables even impossible to strictly apply input harmonization we believe that these pan-European programs have to move in this direction. Even if, as can be assumed, not all member states agree on a blueprint for a questionnaire or on a set of data collection procedures, Eurostat could propose such a blueprint and develop a set of best practice rules for data collection.[16] Although these documents would not be legally binding their existence will lead to them being adopted by many countries because doing so will save time and money. Secondly, to be able to assess data quality in more detail all survey documents should be made available. Aside of questionnaires these would ideally include interviewer instructions and data on the data collection process as is common practice in social surveys. Thirdly, the harmonized and integrated datasets distributed by Eurostat should also contain the original country-specific measures at least for variables for which the harmonization process necessarily in a high information loss. The availability of these data would enable researchers to assess the quality of the harmonized measures and it would allow the construction of alternatively harmonized variables.

6. Conclusions

In this section we have dealt with selected problems of data documentation and data access. We have not addressed the data exchange on the international level which has by and large positively developed in Germany. Foreign scientists have a variety of opportunities nowadays to analyze German data. International research and data centres would be a further step for improving cooperation in research and teaching.

16 This strategy has been already applied with respect to the ICT business survey (Eurostat 2007).

We also have only briefly touched the progress which has been made in broadening the bases of empirical research. A number of activities aim at the generation of complex data bases which combine different data types. The typical micro-macro-dataset is only one example of a large variety of new sources for analysis. Empirical data can be combined with literature and publications, survey data can be combined with regional information, media data etc. In order to create these new data bases metadata standards in particular the DDI standard, have to be further developed (see Heus et al. in this volume). New tools enabling to link different metadata basis are necessary. Some of these tools are currently developed in the context of the Preparatory Phase Project of the Council of European Social Science Data Archives (CESSDA). Interoperable metadata bases finally will help to combine datasets from different years and/or different countries thereby enlarging our resources for inter-temporal and comparative research.

References:

Alda, H./Bender, S. and Gartner. H. (2005): The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB). In: Schmollers Jahrbuch 125, 327 - 336.

Ehling, M. (2003): Harmonising Data in Official Statistics: Development, Procedures, and Data Quality. In: Hoffmeyer-Zlotnik, J.H.-P. and Wolf, Ch. (Eds.): Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York, 17 - 31.

Eurostat (2007): Methodological Manual for Statistics on the Information Society. Survey year 2007, v2.0. Luxembourg: Eurostat. http://europa.eu.int/estatref/info/sdds/en/isoc/isoc_metmanual_2007.pdf (accessed 30/09/2008).

Granda, P. and Wolf, Ch. (2009): Harmonizing Survey Data. In: Harkness, J. et al. (Eds.): Multination, Multicultural and Multiregional Survey Methods. New York (in print).

Janßen, A. and Bohr, J. (2006): Microdata Information System MISSY. Iassist Quaterly 30, 5 - 11.

Lüttinger, P./Köhne-Finster, S. and Urban, J. (2007): Ergebnisse der dritten Befragung von Nutzern der Mikrozensus Scientific Use Files. GESIS Methodenbericht Nr. 1/2007, Mannheim.

Lüttinger, P./Schimpl-Neimanns, B./Wirth, H. and Papastefanou, G. (2004): The German Microdata Lab at ZUMA: Services provided to the scientific community. In: Schmollers Jahrbuch 124, 455 - 467.

Müller, W./Blien, U. and Wirth, H. (1995): Identification Risks of Microdata. Evidence from experimental studies. In: Sociological Methods & Research 24, 131 - 157.

OECD (2007): Education at a Glance 2007. OECD Indicators. Paris.

Pollak, R./Wirth, H./Weiss, F./Bauer, G. and Müller, W. (2009): Issues in the Comparative Measurement of the Supervisory Function using the examples of the ESS and the EU-LFS. In: Pfau-Effinger, B./Magdalenić, S.S. and Wolf, Ch. (Eds.): International vergleichende Sozialforschung: Ansätze und Messkonzepte unter den Bedingungen der Globalisierung. Wiesbaden.

Schneider, H. and Wolf, Ch. (2008): Die Datenservicezentren als Teil der informationellen Infrastruktur. In: Rolf, G./Zwick, M. and Wagner, G.G. (Eds.): Fortschritte der informationellen Infrastruktur in Deutschland. Baden-Baden.

Schneider, S.L. (2008): Suggestions for the cross-national measurement of educational attainment: refining the ISCED-97 and improving data collection and coding procedures. In: Schneider, S.L. (Ed.): The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries. Mannheim, 311 – 330.

Statistisches Bundesamt, GESIS, and WZB (Eds.) (2008): Datenreport 2008. Ein Sozialbericht für die Bundesrepublik Deutschland. Bonn: Bundeszentrale für politische Bildung.

Wirth, H. (2008): Microdata access and confidentiality issues in Germany. Presentation at the meeting “Census Microdata: findings and futures”, University of Manchester, 1 - 3 September 2008.

Wirth, H./Zühlke, S. and Christians, H. (2005): Der Mikrozensus als Datenbasis für die Regionalforschung. In: Grözinger, G. and Matiaske, W. (Eds.): Deutschland regional. Sozialwissenschaftlichen Daten im Forschungsverbund. München, 65-80.

Wissenschaftsrat (2007): Recommendations on the Role and Future Development of Governmental Research Agencies with R&D Activities. http://www.wissenschaftsrat.de/texte/7854-07.pdf.