German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper No. 46

Providing data on the European level

Peter Elias

November 2008

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.

Towards an improved research infrastructure for the social sciences: future demands and needs for action

Providing data on the European level

Peter Elias

November 2008

1. Introduction

This paper reviews the potential demand for and the provision of European data for social scientific research. The concept of *data provision* is defined broadly, covering the ease with which specific types of data can be discovered, interpreted, readily understood and accessed by researchers.

The paper is structured in the following way. The next section addresses the issue of why researchers need European (as opposed to national) data resources. This leads in to a short section discussing the potential demand for data at the European level. The main section focuses on the nature of various data resources. Finally, the paper concludes with an assessment of the need for new and/or improved data infrastructures and suggests where efforts could be focussed to realise such needs.

2. Why do we need data at the European level?

There are two main reasons for supporting the development of Europe-wide data infrastructures. The first relates to the need to inform social and economic policies which are pan-European in design or operation. As the European Union continues to integrate its economic and social structures, there is a need to understand how such integration operates across the EU, and to identify both strengths and weaknesses in policy implementation. It is primarily for this reason that the European Union, through its statistical agency (Eurostat), coordinates the production and collection of census, survey and administrative data across the EU. The second need for European data relates more to the nature of research in the social sciences which, for the most part, cannot make use of randomised and controlled experiments that typify research in the physical sciences and must rely more on variations across groups and through time to investigate causality. Europe provides wide cultural diversity not simply in the obvious dimensions (language, politics, legal systems, *etc.*) but also across more difficult to measure traits such as cultural values, traditions, beliefs. To the researcher this provides variations that help inform the research process. 'Europe' thus affords the research environment that the physical scientists would otherwise harness in the laboratory.

3. What kinds of data do we need for research in the social sciences at the European level?

European-level research has the same basic needs for data as at the national level. However, the very nature of the European Union dictates that there will be specific research interests which may not have any national counterpart. For example, research on cross-national migration within the EU or across its external borders, not just in terms of demographic change but also the assimilation of and attitudes towards migrant groups, requires large comparable micro-level data from all member states. Comparative research on poverty and social exclusion; research to improve our understanding of the political drivers of change; and studies of healthcare delivery across the European Union need this transnational approach. Equally, research to inform our knowledge about processes of economic growth and decline within the European context (*e.g.* trans-national investment, impact studies for the

location of large-scale infrastructures, economic stability within the eurozone) requires a specific Europe-wide focus whilst drawing upon what are essentially national data resources.

4. Pan-European data resources

This section illustrates the available data resources designed to facilitate European research. No distinction is made here between data resources which are purpose built for comparative research at the European level (input harmonised) and those which have arisen as research groups have attempted to meld a number of separate resources into a pan European resource (output harmonised).

To document the variety of data resources that are available, the following typology is adopted:

- *cross-sectional micro resources* – information which is descriptive of a unit of observation at a single point in time. Cross-sectional micro data observations may be repeated in order to monitor change at the macro-level;
- *longitudinal micro data resources* – information which describes the evolution of a unit of observation (*e.g.* a person, a family, an organisation) through time. Such data resources are powerful instruments in the study of cause and effect;
- *macro data banks* – derived from cross-sectional survey or administrative data sources, 'databanks' are repositories of tabulated data, usually providing a wide range of social and economic indicators.

Macro data banks are not covered in detail in this paper. While they constitute important resources for a variety of research interests, access to these resources and their use is relatively easy and uncontroversial[1]. However, for most research purposes, researchers want access to the underlying microdata resources from which the statistical indicators in macro data banks are constructed.

Other typologies are useful, for example the distinction between administrative and transactions data – the former referring to data generated as a by-product of an administrative process (registration for social security benefits) or from a transaction (*e.g.* purchase or sale of goods or services). Important data with potential for research are generated in the commercial sector (not just expenditure data but geo-demographic 'lifestyle' data) and both the commercial sector and the broader academic research communities can benefit from a mutual collaboration. Reference to such data types is made in the concluding section.

[1] As an example of a research resource dedicated to providing access to and information about a wide variety of macro databanks, see ESDS International.

4.1 Cross-sectional microdata resources

4.1.1 Resources available via Eurostat

Since the late 1960s the European Union (formerly the European Community) has sought to develop comparable microdata resources in order to measure and progress social, political and economic integration. These efforts have given rise to a number of major data resources. However, access to these resources has, until recently, been severely restricted.

Cross-sectional micro data collected by Eurostat from National Statistical Offices across the EU include:

- EU Labour Force Survey (EU LFS)
- Community Innovation Survey (EU CIS)
- Structure of Earnings Survey (EU SES)
- Statistics on Income and Living Conditions (EU SILC)

Brief details of each of these sources are shown in the boxes below. Further information can be gained by following the hyperlinks in each box.

Figure 1: Cross-sectional microdata resources available through Eurostat

EU Labour Force Survey

The EU-LFS is the longest running EU-wide statistical survey. Conducted by National Statistical Offices in member states, the LFS has, since 1992, had a common output requirement in terms of the employment-related information it provides on individuals and households. Data currently available covers the period 1983-2006. In Spring 2002 the total sample size was approximately 1.5m persons. Data are available as anonymised micro records.

For further information on access conditions, see: EU-LFS

EU Structure of Earnings Survey

The EU SES is a large enterprise-based sample survey designed to provide accurate and harmonised data on earnings across the EU. The survey was held in 1995, 1999, 2002 and 2006. Results for 1995 are not comparable with later years.

Data collected includes earnings, age, gender, occupation, sector, hours worked, education and training for employees of enterprise with 10+ employees. The latest data available for research purposes is the 2002 survey.

Access to SES data is through the SAFE data access centres in Luxembourg (see EU-SES).

EU Community Innovation Survey

Community Innovation Statistics are produced in all 27 EU countries, 3 EFTA countries and candidate countries. Data are collected on a four year cycle. The first (pilot) survey was held in 1993, the second survey held in 1997/98 and the third survey in 2000/01. The fourth survey, conducted in 2006 with a reference year of 2004 will be available shortly. Anonymised microdata are available via CD-ROM. Non-anonymised data are available through the SAFE data access facility in Luxembourg (see EU-CIS).

The CIS provides information on the characteristics of innovation at the enterprise level.

EU Statistics on Income and Living Conditions

The EU SILC was designed as a successor to the European Community Household Panel which ran from 1994 to 2001. The first release of EU SILC was in 2004, with a 2003 reference year.

Anonymised microdata from 2004 and 2005 are available via CD-ROM.

The EU SILC contains a longitudinal element covering a four year period. The first longitudinal database was made available late in 2007.

For details, see EU-SILC

4.1.2 Resources available via other data providers

4.1.2.1 Luxembourg Income Study (LIS)

LIS began in 1983 under the joint sponsorship of the government of Luxembourg and the Centre for Population, Poverty and Policy Studies (CEPS), becoming an independent body in 2001. The LIS archive contains two databases, the Luxembourg Income Study database and the Luxembourg Wealth Study (LWS), covering cross-national microdata sets on incomes, wealth, employment and demography. The LIS database contains nearly 200 datasets organised in six time periods (waves) spanning the years from 1968 to 2005[2].

With the exceptions of Portugal and Romania for Wave VI (around 2004) and Slovenia for Wave V (around 2000), income microdata are available for all EU countries, North America, Australia, Israel and Taiwan. The newer LWS database (released in December 2007) contains 13 wealth datasets from 10 countries[3].

No direct access to the micro datasets is permitted. Registered users submit syntax (SAS, SPSS, and STATA) which LIS staff run on their behalf. Planned developments is the period 2008-2013 include a web-based user interface for syntax submission, storage of and access to prior programs and an online tabulator. For further information about LIS, follow this link: LIS/LWS

4.1.2.2 Council of European Social Science Data Archives (CESSDA)

CESSDA is a network which promotes the acquisition, archiving and distribution of electronic data. The networks now extends to 20+ countries across Europe, providing access to and delivering over 50,000 data collections per annum and acquiring over 1,000 data collections each year. The CESSDA portal provides easy access to the catalogues of member organisations.

Via its multilingual search interface, CESSDA guides enquirers to appropriate datasets at specific data archives[4]. Enquirers can browse datasets by topic and by keywords before linking to specific archive websites to determine access conditions.

---

2 Microdata held by Eurostat are confidential data about individual statistical units. The release of these data to *bona fide* researchers is governed by Commission Regulations EU Nos. 83/2002, 1104/2006 and 1000/2007 which implement Council Regulation (EU) No. 322/97. Article 17 allows the EU to grant access to data it has collected from national statistical authorities if the national statistical authority gives explicit permission for such use.

3 Austria, Canada, Cyprus, Finland, Germany, Italy, Norway, Sweden, UK, US.

4 CESSDA currently facilitates keyword searches across the following data publishers: UK Data Archive, SSD (Sweden), SIDOS (Switzerland), NSD (Norway), GSDB (Greece), GESIS-ZA (Germany, FSD (Finland), DDA (Denmark), DANS (Netherlands), ADPSS-Sociodata (Italy), ADP (Slovenia)

In 2007, CESSDA acquired FP7 Preparatory Phase funding to facilitate a significant upgrade in its functionality. This three year phase will result in a plan to facilitate and coordinate national funding to provide a designated integrated European research infrastructure. CESSDA also provides access gateways to other important EU-wide data resources, including the European Social Survey, the Eurobarometers, the International Social Survey Programme and the European Values Study (see below for further details about these sources). It also provides researchers with an access point to important data collections outside of Europe, including the major resources avaiable at ICPSR in the US (add hyperlink).

For further information about CESSDA, follow this link: CESSDA.

4.1.2.3 Integrated Public Use Microdata Series-International (IPUMS-I)

IPUMS-I is a project largely funded by the US National Science, based at the University of Minnesota, dedicated to the collection and distribution of census data from around the world.

To date 35 countries have donated microdata from 111 censuses, totalling 263 million person records. The eight European countries which have so far contributed to the IPUMS-I database are Austria, France, Greece, Hungary, Netherlands, Romania, Spain and the United Kingdom. Census data for Slovenia will be available in 2009. Plans are also underway for the addition of censuses from the Czech Republic, Germany, Ireland, Italy, Switzerland and Turkey. The IPUMS-I website maintains good metadata documentation standards that allow users to appreciate differences in the ways in which censuses have been carried out, difference in definition of key variables, *etc.*

Census data are freely available to registered users at IPUMS-I

4.1.2.4 European Social Survey (ESS)

The ESS is an academically directed social survey designed to provide information on the attitudes, beliefs and behaviours of Europe's changing population. Now in its fourth round, the ESS has been mapping long term attitudinal and behavioural changes in European society. Over 30 European countries now participate in the survey, with sample sizes ranging from 1,000 to 2,000 persons in each country[5].

A major strength of the ESS is its attention to methodological weaknesses in the generation and use of cross-national comparative data. Particular emphasis is placed on the interpretation of key concepts in the survey research instruments and their translation into different linguistic and cultural contexts.

[5] The minimum number of achieved interviews is set at 2,000 persons, except in countries with a population of less than 2 million, where the minimum number is 1,000.

4.1.2.5 Eurobarometer

The Eurobarometer surveys were established in 1973, designed to provide the European Commission with data on social trends, values and public opinion generally, helping in the preparation of EU wide policy and to inform the evaluation of its work. Surveys are conducted annually, with each survey covering approximately 1,000 face-to-face interviews[6] in each EU country.

Eurobarometer microdata are available from a variety of sources, including the Inter University Consortium for Political and Social Research at the University of Michigan and the German Zentralarchiv (GESIS-ZA). Links to these sources can be made through the CESSDA network (see above).

4.1.2.6 International Social Survey Programme (ISSP)

Since 1983 the ISSP has promoted cross-national collaboration in the creation of research instruments and methods to generate a wide variety of data about social, economic and political change, values, beliefs and motivations. While individual country samples are fairly small, the ISSP devotes considerable resources to ensuring good comparability between countries.

Further information about the ISSP is available at the following link: ISSP. The data are available via the CESSDA network.

4.1.2.7 European Values Study

The European Values Surveys (and the companion World Values Surveys) are designed to enable a cross-national, cross-cultural comparison of values and norms on a wide variety of topics and to monitor changes in values and attitudes across the globe. Topics covered include perception of life, family, work, traditional values, personal finances, religion and morale, the economy, politics and society, the environment, allocation of resources, contemporary social issues, national identity, and technology and its impact on society. To date four waves have been conducted in 1981-1984, 1990-1993, 1995-1997, and 1999-2004. Not all of the earlier surveys employed probability sampling procedures. These survey responses have been integrated into one dataset, to facilitate time series analysis.

Further information about the EVS can be found at this link: EVS. The data are available via the CESSDA network.

[6] Variations are Germany (2,000), Luxembourg (600), UK (1,300 of which 300 in Northern Ireland).

4.1.2.8 European Working Conditions Survey (EWCS)

The EWCS series began in 1990-91 and is usually conducted every five years. The survey utilises a face-to-face questionnaire administered to a random sample of employed people (employees and self employed) representatives of the working population in each EU country. The latest survey, held in 2005, covered the EU27 plus Croatia, Turkey, Switzerland and Norway.

The questionnaire covers many aspects of working conditions, including violence, harassment and intimidation at the workplace, management and communication, work-life balance and payment systems.

The EWCS datasets for 1991, 1995, 2000 and 2005 are available from the UK Data Archive (Economic and Social Data Service), as part of the CESSDA network. For further information, see EWCS or EWCS at ESDS.

4.1 Longitudinal microdata resources

4.2.1 European Community Household Panel (ECHP)

The European Community Household Panel (ECHP) is a panel survey in which samples of households and persons were interviewed each year from 1994 to 2001[7]. These interviews covered a wide range of topics concerning living conditions. They included detailed income information, financial situation in a wider sense, working life, housing situate on, social relations, health and biographical information of the interviewed. The total duration of the ECHP was 8 years, running from 1994-2001 (8 waves).

For further information, follow this link: ECHP.

4.2.2 Survey of Health, Ageing and Retirement in Europe (SHARE)

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 30,000 individuals aged 50 or over. Eleven countries have contributed data to the 2004 SHARE baseline study, ranging from Scandinavia (Denmark and Sweden) through Central Europe (Austria, France, Germany, Switzerland, Belgium, and the Netherlands) to the Mediterranean (Spain, Italy and Greece). Further data have been collected in 2005-06 in Israel. Two 'new' EU member states - the Czech Republic and Poland - as well as Ireland have joined SHARE in 2006 and participated in the second wave of data collection in 2006-07. The survey's third wave of data collection will collect detailed retrospective life-histories in sixteen countries in 2008-09, with Slovenia joining in as a new member.

[7] Not all countries of the EU participated in the ECHP in all these years.

For further information, follow this link: SHARE. Is now available via GESIS-ZA at part of the CESSDA network.

5. Summary: future needs for European data infrastructure

Table 1 attempts to summarise briefly this review of available *European* data resources which are likely to be of interest to social scientists. The list covers microdata resources only. Macro databanks, providing indicators of trends and yielding information on country and regional differences across Europe, are useful research resources, but do not provide the flexibility needed for exploring social, economic, demographic processes in depth, nor are they adequate for most scientific modelling purposes. The table also excludes CESSDA, which (amongst other functions currently under development) acts primarily as a networked intermediary organisation. The facility it offers, to search data catalogues in different ways across a range of archives in various countries for specific sources of data, makes CESSDA a powerful tool for data discovery and for comparative research where data permits. CESSDA also provides links to many of the resources shown in Table 1, but it is not, in itself, a producer of pan-European data for research purposes.

Table 1: European microdata requirements: sources and issues relating to access, coverage and quality

Type of microdata required	Source	Data access, coverage, quality issues
Census data (demography and housing)	IPUMS-I	Incomplete coverage across the EU.
Labour force, income data, living conditions	Eurostat (EU-LFS; EU-SILC; EU-SES)	High costs of access, complex bureaucracy, some data to be accessed on site, no record linkage possible.
Values, beliefs, attitudes	Eurobarometer; ISSP; European Values Survey	Relatively small sample sizes.
Social and political behaviour	European Social Survey; ISSP	High quality, but relatively small samples.
Longitudinal individual and household data	Eurostat (ECHP); SHARE	ECHP discontinued after 2001. SHARE has high quality longitudinal data but is still fairly new.
Organisation-based data (structure, pay, working conditions)	Eurostat (EU-CIS); European Working Conditions Survey (EWCS)	Only available at 5 year intervals, no longitudinal measures.

This issues that are raised about sample sizes, data accessibility and/or data quality paint a non-too-inspiring picture of the range and availability of European data resources for research across the social sciences and in related disciplines. Despite the efforts made by individuals, research teams and from some national bodies, the availability, accessibility and quality of these data resources is fairly limited.

There are a number of notable exceptions here, particularly the European Social Survey and the Survey of Health, Ageing and Retirement in Europe, both of which, like CESSDA, have been recognised by the European Strategy Forum for Research Infrastructures (ESFRI) and the European Commission as major research infrastructures in need of further support and development. However, in a number of EU countries and North America, major advances are being made to facilitate a broader social science research agenda which encompasses research in the fields of environmental sciences (climate change, air soil and water pollution, crop modification), medical sciences (genetic expression and human behaviour, spread of contagious diseases, impact of ageing) and engineering (transport systems and congestion, housing design, personal and collective security). This broader agenda has required new types of data structures which are significantly larger than any of the resources currently available, are longitudinal in nature and which can be readily enhanced via linkage to administrative and/or transactional data. Simultaneously, new access procedures have been developed and implemented in a number of countries which take advantage of technical developments to provide better and more secure access to complex and sensitive data sources, as well as facilitating a more 'hands-on' approach to research[8] than has been the case with, say, the Luxembourg Income Study or the Eurostat SAFE access procedures.

Possibly the most disappointing aspect of this review relates to the continued barriers to widespread access by the research community to the purpose-built European statistical databases held by Eurostat. Notwithstanding renewed legislative efforts to improve matters from within Eurostat, access remains slow, costly and restrictive. No remote access is provided by Eurostat, despite the proven technology, the security this approach offers compared with the proliferation of data via physical media, the reduced costs and the convenience it provides to the research community. The costs currently incurred by researchers working on publically-funded research are hardly defensible[9].

This suggests where efforts should be focussed to improve these essential research resources. Four major new initiatives are proposed:

5.1 A new European Household Panel

This should build upon the latest developments in a number of countries, to establish larger and better household panels than has hitherto been the case. The obvious first step here is to determine how certain countries can take the household panels they have under academic direction, and align their activities to facilitate cross-panel analysis. There is nothing new in this approach. Indeed, the demand for cross-national equivalent files that have been based upon the US Panel Study of Income Dynamics, the German Socio-economic Panel and the British Household Panel Survey testifies to the values of such resources. However, the new UK Household Longitudinal Study (Understanding Society), the German Socio-economic Panel and the Swiss Household Panel are candidates for renewed efforts to

8 See for example http://www.norc.org/projects/data+enclave+project.htm and http://www.ons.gov.uk/about/who-we-are/our-services/vml/index.html

9 An example of this is the €8,000 cost for a DVD and CD-Rom(s) containing a set of quarterly/yearly files covering available data in 26 countries and all years from 1983 to 2006.

build bigger, better and more comprehensive household panel studies for a number of European countries than has been possible so far.

5.2 Comparative Birth Cohort Studies

A number of countries[10] have commenced work to develop new and bigger birth cohort studies than have been available previously. The opportunity to exploit the rich variety of data these studies will provide and the disciplines that must combine to make this happen (genetics, psychology, economics, sociology, education) provide a world-class opportunity that Europe should grasp.

5.3 Longitudinal studies of organisations

Comparative longitudinal studies of organisations are required, to provide valuable insights into the ways in which enterprise grow, succeed, prosper and decline in an increasingly risky global business environment. The framework for such a development exists in a number of countries (*e.g*. the Workplace Employee Relations Studies in the UK, the REPONSE surveys in France, the database of organisation data held by the German IAB) could form the core of a proposal to develop such a comparative research resource built upon existing surveys and research expertise.

5.4 Improved access to Eurostat data

Last, but not the least of these proposals, is the need to improve access to data held by Eurostat. In part, the problems of access currently faced by researchers are the responsibilities of the National Statistical Institutes which supply the data to Eurostat and which stipulate conditions for their release. This results in what is termed, the 'lowest common denominator' problem. For example 26 out of 27 countries stipulated that identifying information on individual records (*e.g*. names of individuals, names of organisations) should never be made available to researchers. But good research proceeds by allowing researchers to link between data sources, maximising their utility and facilitating new and important research to be conducted. Concerns about data security can now be addressed via the new forms of control and access that virtual remote access provides. There is a now a pressing need to address these issues and to find innovative solutions to unlock the research potential of these truly European resources that cost the EU taxpayer many millions of Euros to create.

In addition to these specific proposals to develop new or to build on existing research infrastructures at the European level, there is a need to determine the feasibility of promoting access to some less well established types of data within a European context. The two most obvious sources of information here are a*dministrative data* sources and *transactions data*. The former derive from the administration of systems or programmes (*e.g*. social security benefit, school records) and can often be mapped onto other resources to enhance their research potential. As a by-product of systems which are not primarily designed to provide research data, and because they are national in character, potential here may be limited, but further investigation of their research potential is warranted. Transactions data are often

[10] These include the UK (a 2012 birth cohort of up to 60,000 persons), Germany (a proposed national birth cohort beginning in 2011), France (a cohort commencing in 2009), the USA (a cohort commencing from 2008 to 2012) and other cohorts in Ireland, Sweden, *etc*.

held by private sector organisations and relate to the delivery of services or to customer-initiated transactions (*e.g.* mobile phone records, shopping data). Where such companies are providing services across the European Union, the potential to use such information for Europe-wide research purposes becomes feasible. However, companies are likely to restrict access and to limit the nature of research that can be conducted from such sources. Again, some preliminary work needs to be undertaken to investigate the feasibility of using such data as Europe-wide research resources.

All these developments will need to be linked to improved training / promotion in/of cross-national / comparative data analysis. Good access to data can only help promote pan-European research efforts if user communities have the capacity to utilise such data resources.

Acknowledgements

A significant part of the information used in this report has been gleaned from the websites of the various data resources which are described. Where possible, the most appropriate link to these websites is included vie hyperlinked text to allow the reader to browse the original documentation and/or to be provided with more up-to-date information than would otherwise be the case. The copyright of all material so presented is hereby acknowledged.

I would like to thank Kevin Schürer at the University of Essex for helpful comments on an earlier draft.