Schmollers Jahrbuch 130 (2010), 421 – 443
Duncker & Humblot, Berlin

# Documentation

## Building on Progress – Expanding the Research Infrastructure for the Economic, Social, and Behavioral Sciences

By The German Data Forum*

*The German Data Forum (RatSWD) adopted the recommendations documented here at its 25th meeting on June 25, 2010, in Berlin. The recommendations are published together with the underlying expert reports in a two-volume compendium: German Data Forum (RatSWD) (ed.), Building on Progress – Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Opladen: Budrich 2010.*

*In June 2010, the members of the Forum were: Roderich Egeler, President of the Federal Statistical Office; Eckart Hohmann, President of the State Statistical Office of Hesse; Frank Kalter, University of Mannheim; Joachim Möller, Director of the Institute for Employment Research and University of Regensburg; Notburga Ott, Ruhr University Bochum; Susanne Rässler, University of Bamberg; Uwe G. Rehfeld, German Federal Pension Insurance; Ulrich Rendtel, Free University of Berlin; Petra Stanat, Director of the Institute for Educational Progress (IQB) at Humboldt University of Berlin; York Sure, President of GESIS (Leibniz Institute for the Social Sciences) and University of Koblenz-Landau, Gert G. Wagner, German Socio-Economic Panel Study (SOEP) and Berlin University of Technology (TUB); and Joachim Wagner, Leuphana University of Lueneburg.*

## Recommendations

### 1. The Big Picture: Measuring the Societies of Progress

The importance of better data for the behavioral, economic, and social sciences is underscored by recent international political developments. For decades, social progress was judged mainly by measures of economic perfor-

* Rat für Sozial- und Wirtschaftsdaten (RatSWD).

mance; above all, by increases in gross domestic product (GDP). In 2009, the Commission on the Measurement of Economic Performance and Social Progress ("Stiglitz Commission")[1] published its report, which opens with the statement that "what we measure affects what we do." It sought to bring about a change in social and political priorities by advocating that greater emphasis be placed on measures of well-being and of environmental and economic sustainability.

The Stiglitz Commission's recommendations form a backdrop to this report.[2] Recommendation 6 in particular can serve as a unifying theme for our recommendations; we quote it below in full.

> ***Both objective and subjective dimensions of well-being are important***
>
> "Quality of life depends on people's objective conditions and capabilities. Steps should be taken to improve measures of people's health, education, personal activities and environmental conditions. In particular, substantial effort should be devoted to developing and implementing robust, reliable measures of social connections, political voice, and insecurity that can be shown to predict life satisfaction."

In Germany, the Statistical Advisory Committee (*Statistischer Beirat)* made the Stiglitz Commission's report the backbone of its recommendations for the next few years. The Committee writes:

> "Initiatives for the further development of national statistical programs – above all demands for new data – often come from supra- and international institutions: the EU Commission, the European Central Bank, the UN, OECD and the IMF. The Statistical Advisory Committee (*Statistischer Beirat*) believes that valuable key initiatives will come from the Stiglitz Commission and the theme *Beyond GDP* advanced by the European Commission. Official statistics, in cooperation with the scientific community, must react to these initiatives and their system of reporting must develop accordingly."

We want to stress this point in particular: *Beyond GDP* will be a fruitful concept only if it is discussed and shaped collaboratively by government statistical agencies and academic scholars. As the Statistical Advisory Committee wrote:

> "The Federal Statistical Office should take stock of the non-official data which may be available with a view to measuring the multi-dimensional phenomenon of *quality of life.* The development of statistical indicators should be undertaken in cooperation with the scientific community."

---

1 Report by the Commission on the Measurement of Economic Performance and Social Progress, chaired by Joseph E. Stiglitz, Amartya Sen and Jean-Paul Fitoussi, http://www.stiglitz-sen-fitoussi.fr, and Stiglitz, J. / Sen, A. / Fitoussi, J.-P. (2010): Mismeasuring Our Lives: Why GDP Doesn't Add Up. New York.

2 International organizations like the Organisation for Economic Co-operation and Development (OECD) are dealing with similar issues. For example OECD established the "Global Initiative on Data and Research Infrastructure for the Social Sciences (Global Data Initiative)" as part of its "Global Science Forum".

Further, at the 12th German-French Council of Ministers in February 2010, President Sarkozy and Chancellor Merkel agreed on the *Agenda 2020*, which included joint work on new measures of social progress. This again was a clear message that policy-makers are interested now more than ever in sound empirical evidence about a wide range of social and economic trends indicative of human progress or regress.

The following principles and themes are not intended to contribute directly to discussion of the Stiglitz Commission report or the initiative of the German-French Council of Ministers. But they do lay the groundwork for improved measurement of economic performance and social progress.

We strongly believe that recent improvements in survey methods and methods of data analysis hold promise of contributing substantially to improved measurement of social progress.

## 2. Background

These recommendations are based on contributions by approximately one hundred social scientists[3] who were invited by the German Data Forum (RatSWD) to write advisory reports on key research issues and future infrastructure needs within their areas of expertise; their reports are published in Part III of the two-volume compendium.[4] The number of experts who have contributed is even larger than it was when the predecessor of this report was published in 2001.[5]

The advisory reports cover a wide range of fields of the behavioral, economic, and social sciences: sub-fields of economics, sociology, psychology, educational science, political science, geoscience, communications, and media research. Some reports focus mainly on substantive issues, some on survey

[3] To avoid long-winded expressions, the term social sciences will be used in the remainder of this report to refer to all the behavioral, economic, educational social sciences, related disciplines.

[4] German Data Forum (RatSWD) (ed.), Building on Progress – Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Opladen: Budrich 2010. All of the expert reports are available as *RatSWD Working papers* as well. See http://www.ratswd.de/eng/publ/workingpapers.html. Some working papers that were not commissioned by the German Data Forum but that are of interest too are available on the homepage of the German Data Forum, especially Working Papers 50, 52, 79, 113, 131, 135, 137, 139, 141, 151 and 153.

[5] Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) (ed.) (2001): Wege zu einer besseren informationellen Infrastruktur. Baden-Baden. For an English translation of the recommendations, see: "Towards an Improved Statistical Infrastructure – Summary Report of the Commission set up by the Federal Ministry of Education and Research (Germany) to Improve the Statistical Infrastructure in Cooperation with the Scientific Community and Official Statistics." *Schmollers Jahrbuch* 121 (3), 443 – 468.

methodology and issues of data linkage, some on ethical and legal issues, some on quality standards. Most contributors work for German academic or governmental organizations, but important reports were also received from individuals in the private sector and from European and American academics. All had a focus on German infrastructural needs, but German as well as international contributors emphasized the importance of international collaborative and comparative research. All reports have been repeatedly peer reviewed; they have been discussed and amended at successive meetings and in working groups organized by the German Data Forum (RatSWD).

We first set out some *guiding principles* underlying the recommendations. The core of the recommendations is structured around a set of *principles* and *specific recommendations* regarding infrastructure for the social sciences.

Research in the fields of public health and social medicine is not reviewed. These are clearly such important and distinct fields that they require their own major reviews.

## 3. Principles Guiding the Recommendations

*Evidence-based research to address the major issues confronting humankind*

The social sciences can and should provide *evidence-based research* to address many of the major issues confronting humankind: for example, turbulent financial markets, climate change, population growth, water shortages, AIDS, and poverty. In addressing some of these issues, social scientists in Germany need to cooperate with physical and biological scientists, with scholars in the humanities, and also with the *international community* of scientists and social scientists.

*Competition and research entrepreneurs*

In making recommendations about the future of research funding and research infrastructure, we recognize the importance of competition and research entrepreneurs. This may seem an unusual perspective. In many countries, including Germany, there is a tradition of centralizing research funding and infrastructure decisions. In our view, this is suboptimal. Science and the social sciences thrive on competition – competition of theory and ideas, competition of methods, and competition of infrastructures.

Public funding of research infrastructure is certainly needed because research findings and research infrastructure are public goods and would be undersupplied in a free market.[6] But decisions should not be made in a centra-

[6] See also UK Data Forum (2009): UK Strategy for Data Resources for Social and Economic Research, RatSWD Working Paper No. 131.

lized, top-down fashion – an approach that has the effect of stifling rather than promoting innovation. The experience of the last few years has demonstrated – notably in the field of empirical educational research – that many fruitful new ideas and initiatives can emerge from a decentralized structure that would almost certainly never have resulted from a "master plan." First of all, in Germany the National Educational Panel Study (NEPS) and the Panel Analysis of Intimate Relationships and Family Dynamics (pairfam) are worthy of mention. Both are new panel studies with a long time horizon.

The history of Germany's Research Data Centers and Data Service Centers illustrates the same point. All the Research Data Centers and Data Service Centers established in the last six years were the result of independent initiatives intended to meet distinctive research needs. The KVI laid the groundwork by initiating the establishment of the first six Research Data Centers through central funding. All the later ones have been voluntary bottom-up developments without a central impulse. The Federal Ministry of Education and Research (BMBF) provided some project funding for a few of those. What was crucial was the basic concept for the Research Data Centers, and that was developed by the KVI in its 2001 report.

It is true that the German Data Forum (RatSWD) later institutionalized this framework by establishing a Standing Committee of the Research Data Centers and Data Service Centers (*Ständiger Ausschuss Forschungsdaten-Infrastruktur des RatSWD).* This committee helps the centers to work together and put forward common interests, but it does not initiate new centers. Indeed, we believe that the German Data Forum (RatSWD) should not do so. What is necessary is a common framework for new initiatives that aim to raise Germany's social science infrastructure to a higher level.

In this report we take some further steps towards developing a common framework for research infrastructure in the social sciences. In doing so, we bear in mind the increasing opportunities open to German researchers to contribute to European and international databases and projects, as well as to projects in Germany itself. We formulate some principles and highlight a range of concepts and ideas drawn from the advisory reports.[7]

We do not make detailed recommendations about specific research fields or particular infrastructural facilities. This would run counter to our view that innovative research directions and new ideas develop mainly at the grassroots of scientific and statistical communities. The advisory reports underlying these recommendations did include a large number of recommendations for promoting research in specific fields and on specific issues. A few of these recommendations are included in this report as examples, but in general our

[7] The advisory reports are also summarized in the two-volume compendium – see Part II "Executive Summaries".

approach is to make recommendations about institutions and processes in which competition and research entrepreneurship can flourish. Nevertheless, by providing the advisory reports in Part III of the two-volume compendium (see footnote * above), we hope to give research funding bodies some idea about the budgets that may be needed if particular ideas are put forward by "scientific entrepreneurs."

*The important role of younger researchers*

Closely connected to the need for competition and innovation in science is the need to develop and foster excellent young researchers and ensure that they have sufficient influence in the research community for their ideas and research skills to flourish. It is, in general, true that a centralized research environment favors older, well-established researchers. Almost unavoidably, it is they who are appointed to the main decision-making positions. However eminent they are, their decisions may tend to favor well-established research topics and well-established methods. Innovation, on the other hand, is more likely to come from younger and mid-career researchers.

An important aim and principle underlying this report is to enhance the roles, influence, and opportunities of younger and mid-career researchers. They should be encouraged and given incentives to act as research entrepreneurs, competing to attract funding, develop infrastructure, conduct research, and disseminate new hypotheses and findings. They may, however, have occasion to form research networks among themselves, and this should be supported.[8]

The need to encourage younger researchers is particularly clear in the official statistical offices. They need more freedom to improve official statistics by doing research. Further, with more research opportunities available, employment in official statistical offices will become more attractive to innovative post-doctoral researchers. Recommendations along these lines are developed under Theme 2 below, where we also suggest that it would be valuable to form new kinds of partnerships with private-sector data collection agencies for the performance of specific infrastructure tasks.

*Social science requires improved theory and methods, not just more data*

The main focus of this report is necessarily on research infrastructure and databases, but we want to highlight explicitly the importance of further improvements in social science theory and also in statistical and survey methods.

[8] See the editorial in *Science*, April 2, 2010, Vol. 328, 17, and letters in *Science*, August 6, 2010, Vol. 329, 626 – 627.

Social scientists in almost all fields complain about data deficiencies. The usually unstated assumption is that if only they had the right data, they could do the rest. This is self-serving and misleading. Theory and method are also crucial, and new developments in these domains often go hand in hand with availability of new data sources. The advisory reports published in Part III of the two-volume compendium describe exciting new data sources available to social scientists, including data arising from "digitization", geo-referencing, and bio-medical tests. We make some recommendations about linkages between new and increasingly available data sources and potential improvements to social science theory and method.

*Research ethics and data protection are of growing importance*

Most data in the social sciences are of course data on human subjects. This means that principles of research ethics and privacy need to be observed. In Germany the right to privacy is enshrined in the Federal Data Protection Act (BDSG, *Bundesdatenschutzgesetz*) which protects individuals against the release of any information about their personal or material circumstances that could be used to identify them. Principles of research ethics, on the other hand, are not embodied in law but are dealt with by the scientific community through codes of ethics promulgated by their professional associations.

Due to new technological developments, e.g., remote sensing, data protection and research ethics are of growing importance. Two of the themes outlined below reflect this importance.

## 4. Specific Recommendations

In this section, we summarize insights arising from the advisory reports and subsequent discussions within the German Data Forum (RatSWD). We do this by presenting ten themes. Most of them represent general ideas and fairly abstract recommendations. We aim to encourage debate in the scientific and policy-making communities.

### *Theme 1: Building on success: Cooperation between official statistics and academic researchers*

The German Data Forum's (RatSWD) current activities, as well as the present compendium, build on substantial achievements flowing from the 2001 KVI report. A major theme of that report was the need for improved cooperation between academics and the official statistical agencies, particularly in regard to making official datasets available for academic research. Initially, four Research Data Centers and two Data Service Centers were set up to provide academics and other users with access to official data files and with training

and advice on how to use them. The original Research Data Centers are associated with the Federal Statistical Office, the Statistical Offices of the German *Länder*, the Institute for Employment Research (IAB, *Institut für Arbeitsmarkt- und Berufsforschung*) of the Federal Employment Agency (BA, *Bundesagentur für Arbeit*), and the German Pension Insurance (RV, *Deutsche Rentenversicherung*). Since then, nine more Research Data Centers have been founded (June 2010) and, after being reviewed by the German Data Forum (RatSWD), they joined the group of certified Research Data Centers. It is also worth noting that, after their first three years, all the original Research Data Centers and Data Service Centers were formally reviewed and received positive evaluations.

One of the advisory reports provided for this review offered the observation that, as a result of the Research Data Centers, Germany went from the bottom to the very top of the European league as an innovator in enabling scientific use of official data. It has also been suggested that the Research Data Centers have had benefits that were not entirely foreseen, in that civil servants and policy advisors are increasingly using research-based data from Research Data Centers to evaluate existing policy programs and plan future programs. Civil servants have more confidence in academic research findings knowing that they are based on high-quality official data sources and that the researchers have received advice on how to use and interpret the data.

Official data files have also become more readily available for teaching in the higher education sector as a result of the recommendations of the 2001 KVI report. CAMPUS-Files, based on the Research Data Center files, have been created for teaching purposes and are widely used around the country.

It is important to note that the Research Data Centers have made good progress in dealing with a range of privacy and data linkage concerns that loomed large ten years ago. Particular progress has been made in linking employer and employee data. Research Data Centers have also, in some cases, been able to develop procedures for enabling researchers to have remote access to data once they have worked with officials in the relevant agencies and gained experience in using the data.

Partly due to the progress already made, but mainly due to technological and inter-disciplinary advances, new and more complicated issues relating to data protection, privacy, and research ethics keep arising. Some of these issues emerge because of the increasing availability of types of data that most social scientists are not accustomed to handling, including biodata and geodata. Other issues emerge due to the rapidly increasing sophistication of methods of record-linkage and statistical matching. These issues are discussed in more detail under Theme 8 (“Privacy”) and Theme 9 (“Ethical Issues”).

Based on these considerations, it is recommended that work continues towards providing a permanent institutional guarantee for the existing Research Data Centers. In the best-case scenario, Research Data Centers that belong to

the statistical offices and similar institutions should be regulated by law. At present, the costs of Research Data Centers are borne by the agencies that host them, and users are not required to pay fees of any kind. We believe that this is the best way to run the centers because it ensures maximum use of official data. In the event that funding issues arise in public and policy discussions, it is recommended that cost-sharing and user-pays models be investigated.

It is recommended that methods of obtaining access to a number of important databases that are still de facto inaccessible to researchers be investigated. Examples include criminal statistics and data on young men collected through the military draft system.

In particular, it is recommended that methods of permitting remote data access to Research Data Center files continue to be investigated.

It is recommended that the microdata of the 2011 Census – the first Census in almost 30 years – should be accessible and analyzed in-depth by means of concerted efforts on the part of the scientific community and funding agencies for academic research.

It is recommended that peer review processes be established and sufficient resources allocated to provide "total quality management" also of the data produced by government research institutes (*Ressortforschungseinrichtungen*).

We are in favor of a coordinated and streamlined process. We take a critical view, however, of the current trend towards increasing numbers of evaluations: this is neither efficient nor beneficial to the scientific content.

It is recommended that data providers in Germany collaborate more closely with the European Union's statistical agency, Eurostat.

### *Theme 2: Inter-sector cooperation: cooperation between academics, the government sector, and the private sector*

A major theme of the 2001 KVI report was the need for greater cooperation and collaboration among academic social scientists, official statistical agencies, and government research institutes (*Ressortforschungseinrichtungen*). Since then, it has become clear that in many areas of data collection and analysis, official institutes and academic organizations can form effective partnerships. Such partnerships would be strengthened if younger researchers in both sets of institutions were permitted more independent roles.

Much remains to be done. Academic research teams and official statistical agencies and research institutes probably still do not always realize how much they have to gain from collaboration. But each side must pay a price.

Academics need to understand and respect the social, political, and accountability environments in which official agencies operate. The official agencies (including the ministries and parliaments behind them), for their part, need to

be willing to give up monopoly roles in deciding what specific data to collect and disseminate.

A strong case can be made that the improved level of cooperation that has been seen in recent years between academic social scientists and official statistical agencies and authorities should now be extended to include the private sector as well. Many large social and economic datasets, especially surveys, are collected by private-sector agencies. Since these agencies operate in a competitive market, they need a reasonably steady and secure flow of work in order to be able to make the investments required to maintain high-quality standards in data collection and documentation. Public-private partnerships may be desirable for initiating, attracting funding for, and continuing long-term survey-based projects. The UK's Survey Resources Network has experience in these ventures and may be able to offer useful guidance. Last but not least, a permanent flow of sufficient amounts of work is necessary to ensure competition between private fieldwork firms.

There are many opportunities for methodological investigations carried out in cooperation among academics and government and private-sector survey agencies. One clear example is investigation of the advantages, disadvantages, and possible biases of mixed-mode surveys. Mixed-mode surveys, which are more and more widely used, involve collecting data using a variety of methods, for example, personal interviews, telephone, mail, and Internet. In practice, respondents are commonly offered a choice of method, and the choice they make may affect the evidence they report.

Leaving aside cooperative ventures with public sector and academic clients, it is clear that private sector fieldwork agencies already collect a vast amount of market research data of great potential value to academic researchers.

The potential of market research data for secondary analysis lies mostly in the fields of consumption patterns and media usage. The German market research industry is huge – it has an annual turnover of more than two billion euros – and over 90 percent of its research is quantitative. However, samples are often highly specialized; telephone interviewing is the most common mode of data collection; and data documentation standards are not as high as academic social scientists would wish. However, secondary data analyses seem to be worthwhile – last but not least as a kind of quality control for these data. Clearly, too, the commercial clients for whom data are collected would have to give permission for secondary analysis. The data would have to be anonymized not only to protect individuals, but also to protect commercially sensitive information about products.

In addition, transaction data (e.g., about purchasing behavior) that is generated by commercial firms can be of interest for scientific research. In this case, anonymization is extremely important. The German Data Forum (RatSWD) makes no specific recommendation about this issue beyond the view that recog-

nition of market research data and transaction data merits consideration in the scientific and statistical communities.

### *Theme 3: The international dimension*

The main focus of the detailed advisory reports contained in the two-volume compendium is of course on German social science infrastructure and research needs, but the international dimension is critical too. Plainly, many of the problems with which social scientists as well as policy-makers deal transcend national borders; for example, turbulence in financial markets, climate change, and movements of immigrants and refugees. Furthermore, international comparative research is an important *method of learning*. Similar countries face similar issues, but have developed diverse and more or less satisfactory policy responses. To do valuable international comparative research, researchers usually need to work with skilled foreign colleagues.

International data collected by the EU and other supra-national organizations have important strengths but also important limitations. The data are at least partly "harmonized" and cross-nationally comparable. Generally, however, data coverage is restricted to policy fields for which international organizations have substantial responsibility. Data are much sparser in areas that are still mainly a national-level responsibility. Furthermore, the needs of policy-makers, for whom the data are collected, do not exactly match the needs of scientists.

For example, policy-makers require up-to-date information, whereas scientists give higher priority to accuracy. Policy-makers are often satisfied with use of administrative and aggregate data and accept "output harmonization," whereas scientists favor the collection of micro-level survey data and prefer "input harmonization," that is, data collection instruments that are the same in each country.

We include some recommendations regarding international cooperation, which still raises some difficult problems for German researchers, in part because of legal restrictions on data sharing. Indeed we recommend that a working group be set up by the German Data Forum (RatSWD) to find ways of making German official statistics available to reliable foreign research institutes.

There are several cooperative European ventures that shall be discussed in an open and constructive manner. These include a new European household panel survey under academic direction, Europe-wide studies of birth and other age cohorts, and a Europe-wide longitudinal study of firms. It would also be of great benefit to comparative European research if access to micro-level datasets held by Eurostat could be improved. Ideally, these data would be made available by remote access, with appropriate safeguards to ensure data security.

It is noted that, following a British initiative, an International Data Forum (IDF) has been proposed. Along the lines of the UK Data Forum and the German Data Forum (RatSWD), this body would aim to bring together academic researchers and official statistical institutes, including international organizations like Eurostat. The plan is currently being developed via an Expert Group set up under the auspices of the OECD. It is recommended that Germany participate in this and related initiatives through the German Data Forum (RatSWD) and possibly other bodies.

Finally, it is clear that the academic data providers are not very well organized at the international and supra-national level. Notable exceptions are international survey programs like the European Social Survey (ESS) and the Survey of Health, Ageing and Retirement in Europe (SHARE), and networks of archives like the Council of European Social Science Data Archives (CESSDA), "Data Without Boundaries," and the "Committee on Data for Science and Technology (CODATA)." It is recommended that the academic sector consider setting up an independent organization to represent its interests at the European and worldwide levels. This academic organization would be one of the partners in the international bodies that are likely to be established following the OECD initiative.

***Theme 4: Data on organizations and "contexts"***

It is clear that, since the 2001 KVI report, in Germany a great deal of progress has been made in improving academic researchers' access to firm-level data; that is, to data on employers and employees. These are high-quality data mainly collected in official surveys; firms are required to respond and to respond accurately. Most of the official collection agencies now deposit their data in Research Data Centers. Progress has been made on issues of data linkage, while protecting confidentiality, with the result that it is now often possible for researchers to link data from successive official surveys of the same firm. It is not, however, at present legally possible to link surveys of German firms to international datasets. This would be a desirable development, given that many firms now have global reach.

Progress made in improving access to data on business organizations points the way towards what needs to be achieved in relation to the many other organizations and contexts in which people live and work. Individual citizens are typically linked to multiple organizations: firms, schools, universities, hospitals, and of course their households. Linking data on these organizations and contexts with survey data on individuals would be desirable.

At present, then, there are no German datasets that have adequate information on all the organizations in which individuals operate. So, data need to be collected on respondents' roles and activities in multiple organizations, and where possible, linked to data about the organizations themselves. This could

potentially be achieved by (1) adding additional questions about organizational roles to existing large-scale surveys, perhaps including the large sample of the German Microcensus, and also (2) by linking existing survey datasets to organizational surveys.

A very special kind of a new data type is information about historical contexts, which can be linked to time series data or microdata with a longitudinal dimension. The European Social Survey (ESS), for instance, provides such a databank. It is worthwhile to think about a centralized data center of that kind as a service to the community at large.

Data on political and civil society organizations appear to be in particularly short supply. In many Western countries, evidence about political parties – the most important type of political organization – is regularly obtained from national election surveys. Election surveys are also the main source of evidence on mass political participation. We want to note that in Germany, there is no guaranteed funding for election surveys, although a major election project (GLES, *German Longitudinal Election Study*) is currently being undertaken.

Several of the advisory reports prepared for the German Data Forum (RatSWD) discussed detailed practical ways of realizing these possibilities. It is recommended that funding agencies consult these advisory reports when assessing specific applications to conduct organizational research.

### *Theme 5: Making fuller use of existing large-scale datasets by adding special innovation modules and "related studies"*

Many of the advisory reports recommended that fuller use could be made of existing large-scale German datasets by adding special innovation modules, thereby creating greater value for money. Suggestions were made both for *special samples* and for *special types of data* to be collected. In all cases, it was suggested that the particular benefit of adding modules was that the underlying survey could serve as a national benchmark or *reference dataset* against which the new, more specialized data could be assessed.

The availability of a reference dataset enables researchers to obtain a more contextualized understanding of the attitudes and behaviors of specific groups. Conversely, the availability of detailed and in-depth evidence about subsets of the population can strengthen the causal inferences that analysts of the main reference dataset are able to make.

The advisory reports covering international and internal migration document substantial data deficits, which, it is suggested, could be largely overcome by adding special modules to existing longitudinal surveys. It has been pointed out that existing datasets do not allow researchers to track the careers of migrants over long periods. This is particularly a problem in relation to

highly skilled migrants, a group of special interest to policy-makers. Migrant booster samples, added to existing large-scale surveys, would largely overcome the problem.

Reports written by experts in other fields made similar recommendations. For example, it was suggested that data deficits relating to pre-school education and vocational education and competencies could be partly overcome by adding short questionnaire modules to ongoing surveys.

It is more or less conventional in the social sciences to collect exploratory qualitative data – for example, open-ended interviews – to develop hypotheses and lay the basis for quantitative measures prior to embarking on a large-scale quantitative project. It is suggested that this sequence can also sensibly be reversed. Once a quantitative study has been analyzed, individuals or groups that are "typical" of certain subsets can be approached with a view to conducting qualitative case studies. The researcher then knows precisely what he / she has a "case of." Extended or in-depth interviews can then be undertaken to understand the decisions and actions that subjects have taken at particular junctures in their lives, and the values and attitudes underlying their decisions.[9]

A further suggestion is that innovation modules using "experience sampling methods" be added to existing large-scale surveys. Again, the procedure would be to approach purposively selected respondents, representing sub-sets of the main sample, and ask them to record their answers to a brief set of questions (e.g., about their current activities and moods) when a beeper alerts them to do so.

### *Theme 6: Openness to new data sources and methods*

Advisory reports prepared for the German Data Forum (RatSWD) highlighted the potential of several exciting new sources and methods of collecting data. We want to mention some of these sources, but without making specific funding recommendations. We do, however, want to stress that Germany needs to develop funding schemes that are receptive to inter-disciplinary research proposals involving use of these new data sources and data collection methods.

#### Digitization

It is widely recognized that data grid technology ("digitization") is generating massive amounts of new data that are potentially valuable to social scientists. A great deal of data is generated through the use of the Internet, including e-mail and social networking sites, and through the use of cell phones, GPS

[9] It is important to address the privacy and ethical implications of approaching survey respondents for additional interview data. Clearly, they must be asked for explicit consent to link the data sets.

systems, and radio frequency identification devices (RFIDs). To date, social scientists have made limited use of these datasets, partly because it is not clear how to gain access and how to deal with privacy issues. A few initiatives have been undertaken. For example, the networking site Facebook reports that social scientists in all English-speaking countries are analyzing messages posted on the site each day to assess changes in moods and perhaps happiness levels.

However, it seems unlikely that substantial progress will be made until access and privacy issues are solved. The German Data Forum (RatSWD) notes that the UK's Economic and Social Research Council (ESRC) has set up an Administrative Data Liaison Service to deal with these issues by linking academics to producers of administrative data.

### Geodata – the geo-spatial challenge

Most of the data used in the social sciences have a precise location in both space and time. While geodata are used widely in geography and spatial planning, this is generally not the case in the social sciences. Spatial data from various sources, including remote sensing data, can readily be combined via the georeferences of the units under investigation. This makes georeferenced data a valuable resource both for research and for policy advice and evaluation. While administrative spatial base data have been widely available for Germany for a long time, there has been an enormous increase in recent years in the supply of spatial data collected by user communities (e.g., OpenStreetMap) and private data providers (e.g., StreetView). Furthermore, remote sensing data (aerial photos or satellite data) have become more important. These data come from a number of different places scattered across the globe and are provided by different sources, which makes it important to launch geodata infrastructure projects that bring together different geodata sets. It has to be pointed out that data security is of high importance for this type of data; issues of personal rights are particularly sensitive.

Closely related to geodata are data for regions, which can be defined as areas as large as a German *Land* or as small as a village. Regional data have been available for many years and have been used for cross-regional investigations and as context variables in studies investigating the behavior of persons or firms. Access to many datasets at various levels of regional aggregation is straightforward in Germany through the use of cheap CDs / DVDs and the Web.[10] The main challenge is to offer access to geodata in ways that allow easy combination with other data. Both current and older data need to be made available to allow for longitudinal studies. Furthermore, data for individuals, households, and firms should be entered with a direct spatial reference; this is especially important for the forthcoming 2011 Census.

[10] http://www.geoportal.bund.de, http://www.raumbeobachtung.de, http://www.regionalstatis tik.de. [Accessed on: August 7, 2010].

An important recommendation for the future is to intensify collaboration between social science researchers and researchers in institutions in the currently rather segregated areas of geoinformation and information infrastructure. Thus, the German Data Forum (RatSWD) will set up a *working group* on geodata and regional data with a view to bringing the different data providers and users together.

### Biodata: research incorporating the effects of biological and genetic factors on social outcomes

In recent times, greater attention has been paid in the social sciences to biomedical variables, including genetic variables that influence social and economic behaviors. Many opportunities, and some serious risks, exist in this growing research field. Historically, social scientists have received no training in biomedical research and are unlikely to be aware of the possibilities. Certainly, they have little knowledge of appropriate methods of data collection and analysis. It is under discussion whether the German Data Forum (RatSWD) will set up a *working group* with a view to positioning German social scientists to be at the forefront of developments. The group would need to include biologists and medical scientists, as well as social scientists and – equally important – not only data protection specialists but also ethics specialists. In addition, one issue that such a working group would have to address is the difficulty that researchers who are working at the interface of the social and biomedical sciences currently have in attracting serious funding.

A role model for this kind of data collection may be found in the SHARE study, which has already conducted several pilot studies, collecting biomedical data from sub-sets of its European-wide sample. It has been shown that, with adequate briefing, medically untrained interviewers can do a good job of getting high-quality data, without a significant increase in interview refusals and terminations.

### Virtual worlds for macro-social experiments

Advocates of the use of computer-generated "virtual worlds" (such as "Second Life") for social science research believe that they offer the best vehicle for developing and testing theories at a "macro-societal" level. Many of the problems facing humanity are international or threaten whole societies: climate change, nuclear weapons, water shortages, and unstable financial markets, to name just a few. By setting up virtual worlds with humans represented by avatars, it is possible to conduct controlled experiments dealing with problems on this scale. The experiments can be run for long periods, like panel studies, and they can allow for the involvement of unlimited numbers of players. They pose no serious risk to players and avoid the ethical issues that limit many other types of study.

Advocates of macro-social experiments recognize that initial costs are high, but claim that the worlds they create hold the prospect of eventually being self-funding, paid for by the players themselves.

### *Theme 7: Data quality and quality management*

This theme deals with issues relating to (1) the quality of available measurement instruments, and (2) the quality of documentation required to facilitate secondary analysis of existing datasets.

Experts in several areas in their advisory reports made the point that a fairly wide range of measurement instruments were available to them, but that researchers would benefit from guidance in assessing their comparative reliability, validity, and practicality in fieldwork situations. In the advisory reports, it was suggested that something like a *central clearing house* was needed with a mandate to assess and improve standards of measurement. It was noted that the recent founding of the Institute for Educational Progress (IQB, *Institut zur Qualitätsentwicklung im Bildungswesen*) could serve as a model.

The Institute was launched at a time when the poor performance of German students in standardized international tests led to increased concern with measuring learning outcomes. The IQB is measuring the performance of representative samples of students in the 16 German *Länder,* and will also be available to serve as a source of advice on measurement issues

A related but somewhat separate concern mentioned in several advisory reports is the poor quality of documentation provided for many surveys and other datasets that, in principle, are available for secondary analysis. It appeared that the academic sector has much to learn in this respect from the official sector, which generally observes high standards in data collection and documentation.

In thinking about data storage and documentation, a distinction should probably be drawn between two types of academic projects: those that are of interest only to a small group of researchers and those that are of wider interest. A mode of self-archiving (self-documentation) should suffice for the former type, although even here minimum satisfactory uniform standards need to be established. The latter type should be required to meet high professional standards of documentation and archiving (see Theme 10).

To a large extent, improvement of survey data documentation is a matter of adopting high *metadata standards.* These are standards relating to the accurate description of surveys and other large-scale datasets that need to be met when data are archived. Historically, researchers paid little attention to the quality of metadata surrounding their work; archiving was left to archivists. This mindset is changing. There have been rapid advances in the development and implementation of high-quality metadata standards, standards which apply to da-

tasets throughout their life cycle from initial collection through to secondary use, perhaps in conjunction with quite different datasets.

An important source of survey metadata is the information collected about individuals, households, and locations when seeking and interviewing designated respondents. These data, sometimes termed *paradata*, are typically recorded by interviewers and deposited with their survey research agency. The data are valuable for analyzing problems of survey non-response and for assessing the advantages and disadvantages of different data collection modes. Paradata can be used to attempt "continuous quality improvement" in survey research. It is recommended that efforts be made to standardize and improve the quality of paradata collected by public and private-sector survey agencies. The European Statistical System has published a handbook on enhancing data quality through effective use of paradata.

In Germany, the Research Data Centers have taken the lead in trying to improve current standards. Based on their experience, it appears that there are two internationally acceptable sets of metadata standards – the Data Documentation Initiative (DDI) and the Statistical Data and Metadata Exchange (SDMX) Standard – which could be more widely used in Germany. Adoption of these standards requires the establishment of a registry-based IT infrastructure compatible with the industry standard for Web services. This infrastructure can then facilitate the management, exchange, harmonization, and re-use of data and metadata.

We would like to highlight one potential means of improving documentation in particular: the use of a unique identifier for datasets (e.g., a digital object identifier or DOI). Unique identifiers for particular measurement scales (e.g., the different versions of the "Big Five" inventory) could possibly also be helpful (see also Theme 10 below).

The need for high-quality metadata appears even more pressing when recalling that many Internet users who are not themselves scholars are making increased use of these data for their own analyses. Results generated by lay users are especially likely to be skewed or misleading if the strengths and limitations of the data are described inadequately or in jargon a layperson could not be expected to understand.

### *Theme 8: Privacy issues*

This section deals with privacy issues, particularly those that arise due to increasingly sophisticated methods of data linkage. *Record linkage* refers to the possibility of linking up different datasets containing information about the same units (e.g., individuals or firms). Linkages may be made, for example, between different surveys or between survey data and administrative data. Normally, datasets can only be linked if a common identifier is available.

However, linkage can sometimes now be achieved by means of "statistical matching" when datasets do not contain the same identifiers for particular individuals.

When an individual consents to take part in a specific research project, her commitment – and the limits of that commitment – are usually reasonably clear. But what is the situation if researchers then link a file obtained for this specific project to other files about the respondent, which, for example, contain information about her employer, tax files, health, or geographical location? Clearly, such linked data are of immense value to researchers, both in conducting basic scientific research and in providing policy advice. But do the individuals whose data are being linked need to give specific consent prior to each new linkage?

The advisory reports written for the German Data Forum (RatSWD) expressed a wide variety of views on this matter, with one even describing data linkage as contrary to law and rightly so. We believe that these problems could be resolved best by passing legislation that would require researchers to observe a principle of "research confidentiality" (*Forschungsdatengeheimnis*). This legislation, which was recommended by the KVI in 2001, would require that if authorized researchers obtained knowledge of the identity of their research subjects – even by accident – they would be obliged not to reveal the identities under any circumstances. Most important, the act would prevent both police and any other authorities from seizing the data. When pushing forward the issue of "research confidentiality", it will be important to refer to the European legislation.

A further proposal, or perhaps an alternative, discussed in one of the advisory reports, is for data stewards (*Treuhänder*) to be appointed for the purpose of protecting the privacy of research subjects. Data stewards would be responsible for keeping records of the identity of subjects and would only pass data on to researchers for analysis with the identifying information removed. In Germany, data stewards have recently been used by the official statistical agencies when data linkage exercises have been undertaken. If their use were to be extended to the academic community, their relationships with Research Data Centers would need to be worked out in detail.

A more general recommendation given in the reports is that a "National Record Linkage Center" be set up to cover all fields in which record linkage is an issue. This has been proposed in part to avoid the duplication that would occur if each branch of social science made its own separate efforts. The German Data Forum (RatSWD) makes no specific recommendations but believes that the proposal is worth detailed consideration.

***Theme 9: Ethical issues***

This theme deals with two separate sets of ethical issues: the ethics of research using human subjects, and the ethics of scientists in publicizing their results.

Research using human subjects

The need to define and enforce ethical standards in research using human subjects has always been urgent and has become more so in view of the increasing availability of new types of data highlighted in this report: administrative and commercial data, data from the Internet, geodata, and biodata.

In practical terms, Germany does not yet have a detailed set of ethical requirements that protect research subjects and are designed specifically for the social sciences. However, all researchers have to abide by the requirements of the Federal Data Protection Act. Additionally, the main professional associations in sociology and psychology have issued ethical guidelines, but these mainly affect behavior towards peers, rather than towards research subjects.

A review of ethics procedures in the UK and the US was undertaken by an advisory report to see if they offered useful examples for Germany. British procedures appear worth consideration; US procedures are perhaps too heavily geared towards the natural sciences.

In the UK, beginning in 2006, the *Economic and Social Research Council* (ESRC), which is the main funding body for academic research, forced universities whose researchers were seeking funding from ESRC to set up ethics committees. In practice, committees have been put in place in all universities, usually operating at the departmental or faculty level and not always on a university-wide basis. The committees are required to implement six key principles, four of which protect human subjects. Subjects have to be fully informed about the purposes and use of the research in which they are participating; they have the right to be anonymous; the data they provide must remain confidential; participation must be voluntary, and the research must avoid harm to the subjects.

The principle of "avoiding harm" is particularly important in view of the increasing availability of Web data, geodata, and biodata. "Avoiding harm" appears to be a principle of more practical relevance than the principle of "beneficence" that German social scientists, borrowing from the biological sciences, have sometimes incorporated into ethical guidelines.

Above all, given that research is conducted increasingly on the basis of international exchange, and data are exchanged between different countries and national research institutions, it is of growing importance that respondents be able to rely on users to handle their data responsibly. Due to differences in national data security regulations as well as in research ethics standards, this is a difficult task, which, at worst, can hinder research. However, universal data protection rules are desirable, but extremely unlikely. Thus, it is important

that, at a minimum, the scientific and statistical expert communities raise awareness that universal ethical standards are necessary.

Scientific responsibility in publicizing results

A final key set of ethical issues surrounds the responsibility of scientists in publishing and publicizing their results. In a recent editorial in *Science*,[11] it is noted that "bridging science and society" is possible only if scientists behave properly – that is, in accordance with scientific standards. The editorial mentions not just the need to avoid obvious scientific misconduct relating to data fraud or undisclosed conflicts of interest, but also the importance of avoiding "over-interpretation" of scientific results.

It is worth noting that many economists appear to believe that over-interpretation (by simplifying results) is necessary if a scientist wants to reach the general public. The former Federal President of Germany, Mr. Koehler, an economist, appeared to endorse this approach by calling for social scientists to announce "significant" findings without burying important results under too many details.

We believe that it would *not* be wise for social scientists to take this advice, precisely because scientific results often become the subject of contentious public policy debates. Empirical results *can* have the effect of making policy debates more rational, but only if the assumptions and shortcomings of research are communicated honestly. It is a duty of the scientific community to promote this type of honesty.

### *Theme 10: Giving credit where credit is due*

A key principle of these recommendations is *"to give credit where credit isdue"*. This principle[12] should apply to efforts at developing the social science research infrastructure just as much as to academic authorship. In general, valuable new infrastructural initiatives will only be launched if the staff of infrastructures under academic direction, of official statistical agencies – and perhaps of private-sector organizations that collect and provide data as well – feel recognized and rewarded for undertaking this important work. Junior and senior staff of all types of organizations need to be clearly recognized for their important contributions.

Existing academic conventions about "authorship" are not entirely satisfactory, nor are "science metrics" that evaluate the output of researchers, universities, and research institutes. In a recent article in *Nature*[13] it is suggested:

---

11 *Science*, February 19, 2010, Vol. 327, 921.

12 *Nature*, December 17, 2009, Vol 462, 825.

13 *Nature*, March 25, 2010, Vol. 464, 488 – 89.

> "Let's make science metrics more scientific. To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity. … The issue of a unique researcher identification system is one that needs urgent attention."

Sometimes effective partnerships and joint investments by academic research institutes, official statistical agencies, and private fieldwork organizations occur despite seriously inadequate incentives and recognition. However, in order to make such collaborations more than rare events, the "rules of the game" must be changed. The establishment and running of infrastructure like biobanks, social surveys, and *Scientific Use Files* of register data must be rewarded more adequately than at present. This applies to official statistics, public administrations, private organizations, and the sciences and humanities more generally. The German Data Forum (RatSWD) sees itself as one of the key players in promoting discussion and proposing effective steps on this issue. Here we want to mention two instruments that *might* help to ensure that credit is given where it is due.

First, the establishment of a system of persistent identification of datasets (like the DOI system) would not only allow easier access to data, but also make datasets more visible and citable, and thereby enable the authors / compilers of the data to be clearly recognized. Even particular measurement "devices" (e.g., specific scales for the "Big Five" inventory) might be identified and citable by unique identifiers. A digital object identifier makes it easier to see the links between a scholarly article, the relevant datasets, and the authors / compilers of the datasets. There are already some organizations that have assigned DOIs to datasets (e.g., CrossRef and DataCite).

Second, the issue of a unique researcher identification system is equally important and needs urgent attention. The recent launch of Open Researcher Contributor ID (ORCID) looks particularly promising. The use of a unique researcher ID makes the scientific contributions of each individual researcher who works on a dataset clearly visible.

## 5. Concluding Remarks

In Germany, there are several organizations for funding scientific research. Some policy-makers, government officials, and senior researchers believe that a more centralized organization would do better, but we, the German Data Forum (RatSWD), disagree. Competition opens up more space for new ideas than would be available in a centralized system.

Even though we do not support centralized organization of research, we nevertheless recognize an increasing need to provide long-term funding to establish and run large-scale social science infrastructure. It is clear that both the academic community and those involved in administering Germany's statisti-

cal system are thinking more than ever before about how to reshape and fund their services. So, for example, the German Council of Sciences and Humanities (WR, *Wissenschaftsrat*), and Germany's Joint Science Conference (GWK, *Gemeinsame WissenschaftsKommission*) have working groups underway that are considering matters of research infrastructure.[14] The discussions in these working groups have already made obvious that not only Research Data Centers and data archives but also more and more libraries – university and research institute libraries as well as centralized specialist libraries (*Fachbibliotheken*) – are an important part of the research infrastructure, providing crucial data documentation and access services. Even the Federal Archive (*Bundesachiv*) could play a certain role. Nothing is settled yet. However, it is time to find a new and appropriate division of labor among these institutions.

Thoughtful formulation of key issues and especially the detection of shortcomings and difficulties is itself an important step. Many approaches will no doubt be considered, but in our view it is preferable to develop *principles* for funding and managing research infrastructure, rather than to attempt the almost impossible task of formulating a *master plan.*

The German Data Forum (RatSWD) is itself neither a research organization nor a funding organization. It exists to offer advice on research and data issues. This places it in an ideal position to moderate discussions and help find the most appropriate funding arrangements for the social sciences.[15]

[14] These are (in 2010) the "Research Infrastructure Coordination Group (*Koordinierungsgruppe Forschungsinfrastruktur*)" and the "Working Group on a Research Infrastructure for the Social Sciences and Humanities (*Arbeitsgruppe Infrastruktur für sozial- und geisteswissenschaftliche Forschung*)" of the German Council of Science and Humanities (WR, *Wissenschaftsrat*) as well as the "Commission on the Future of Information Infrastructure (KII, *Kommission Zukunft der Informationsinfrastruktur*)" of the Joint Science Conference by the Federal and Länder Governments (GWK, *Gemeinsame Wissenschaftskonferenz des Bundes und der Länder*).

[15] See also the "Science-Policy Statement on the Status and Future Development of the German Data Forum (RatSWD)" by the German Council of Science and Humanities (WR, *Wissenschaftsrat*). *Schmollers Jahrbuch* 130 (2), 269 – 277.