Why Does Japan Need a Data Archive for the Social Sciences?

Wolfgang JAGODZINSKI

This paper tries to explain the functions and the advantages of a data archive for the Social Sciences. It not only provides the scientific community with useful data sets and information, its tasks are much, much broader. It helps to establish a culture of data-sharing. I think that the ZA (Zentralarchiv für Empirische Sozialforschung) in Cologne, which is primarily an archive of survey data, fulfills most of these functions. I will therefore repeatedly refer to the ZA in order to illustrate my points.

Providing Empirical Information

The first and most obvious advantage of a data archive is the provision of the scientific community and a broader public with empirical information. This does not mean that we engage ourselves in all current political debates and inform the public about the most recent developments. On the contrary, in these debates data archives will and should play a minor role. They cannot compete with the commercial institutes and the primary investigators in offering the latest news and information about the most recent developments for the simple reason that an archive has no immediate and direct access to the surveys. Usually, the data are delivered to the archive after the heated political debate is over, after the voter has decided, or after the issue is settled. An archive’s role is rather that of a collective memory. It can enter data and findings from earlier studies into the debate, it can point to ongoing theoretical discussions or to methodological problems. It should not come up with the most recent election forecast in a campaign, but it might publish an essay on the personalization of politics on the basis of a secondary analysis of its own data. It may also explain the statistical problems of confidence intervals or the methodological problems of voting forecast to a broader audience or the journalists who often are unfamiliar with statistical problems.

Storing Information from the Past

The notion of a collective memory already implies that an archive also stores information from and about the past. The older a data archive becomes the better it can perform this task. In my view, the long-term change of societies belongs to the most fascinating topics in our discipline: Regarding our history, how have Japan and Germany managed to become stable democracies after World War II? Why have they succeeded while many other countries failed? To what extent was the change of institutions accompanied by a change in values and attitudes? In order to answer these questions we need a long sequence of value and attitude surveys with a starting point in the forties or early fifties. An important task of a data archive is, as my predecessor Erwin K. Scheuch has phrased it, continuous observation. Surveys, however, which are not delivered to an archive, are usually short-lived. They tend to disappear with the researcher who has conducted the survey. The data may still be physically present on a shelf in a storeroom but it will require enormous investments and efforts to process them and to bring them into an adequate format. And nobody will make this investment. There may be institutions like the Institute for Mathematics and Statistics in Tokyo which continuously replicate earlier surveys. In my view, their surveys on the Japanese National character belong to the most important sequences of surveys in the world. However, a single study will

Key words: data archive, Zentralarchiv für Empirische Sozialforschung, social science surveys.

**Professor and Dean, Faculty of Economics, Management and Social Sciences, and Director, Central Archive for Empirical Social Research, and Institute for Applied Social Research, The University of Cologne, Germany
necessarily have a limited scope. An archive will have a much broader collection of data. And it cannot restrict the data access.

Survey data usually are lost if they are not appropriately stored. Occasionally, we acquire older data sets which still can be processed with reasonable investments. The ZA was founded in 1960. The managerial director of the archive, Ekkehard Mochmann and my predecessor Erwin K. Scheuch have been successful in acquiring data sets from the fifties. Probably the best known older data set in the ZA is the Reigrotzki study from 1953. In many empirical analyses this study serves as a baseline for comparisons with later developments. Recently, the ZA in co-operation with Professor Rattinger and the Roper Center have been able to rescue a larger number of USIA studies from the early years of the FRG. Surveys are an important but not the only source of quantitative data. The historical department of the ZA, for example, offers data on the composition of the German Reichstag. It also possesses economic time series and aggregated election data. In short, an archive can offer useful information for the social sciences. And it can fulfill its function better the more it is accepted by the scientific community and the older it is.

Maintaining a Staff of Social Scientists

These activities do not only require statistical and technical skills but also theoretical knowledge. The ZA has a number of highly talented social scientists in house. It is one of the most successful institutes in Germany as far as the promotion of academic careers is concerned, particularly in the field of political sociology. The top academic election researchers in Germany were formerly assistants at the ZA. If you are hired in the university department of the ZA you have fairly good chances for later becoming a university professor. However, the social sciences are rapidly specializing and the archive cannot be competent in all fields of a discipline. It therefore becomes increasingly important to co-operate with external experts in those cases where we analyze the data ourselves and publish the results.

Providing High Quality Data Efficiently

Scientific publications and essays are important. However, they are complementary to another activity which is a service for the scientific community in Germany and all over the world: We have to provide researchers as quickly and efficiently as possible with top-caliber data.

The archive was established as an institute for social research. One of its founding fathers, Günther Schmölders, was from management science. One therefore would have expected empirical research in business and management and in sociology or political science to take a similar course of development. This was not the case, however. For several reasons, a much stronger division of labor between private and public institutions emerged in business science. Here, quantitative empirical research is quite often the domain of commercial institutes with little or no contact with the universities. The education in methods and statistics at the universities in turn is rarely based on real data sets. An expert commission on the improvement of the informational infrastructure in Germany (KVI) has clearly diagnosed these deficiencies in business science and economics in a recent report. In my view, they are partly a structural consequence of a completely different incentive structure. In management science the demand for survey information is usually high. Commercial institutes can make a profit by selling the data. As long as this is the case, they have little incentive to deliver them to a data archive. Data sets which are inaccessible for the scientific community, however, are partly immunized against criticism. Conjecture and refutation, however, is, as we have learned from Karl Popper, the driving motor of scientific progress. Thus, it may sometimes be to the advantage of science if the commercial interest in a data set or a class of data sets is low.

Reducing the Cost of Empirical Research

A data archive reduces the costs of empirical research. At first glance, this is an astonishing statement because the archive costs a lot of money. Our yearly budget amounts to about 7 Million DEM, i.e. about 350 million yen. A functioning data archive, however, can offer a large amount of data for empirical tests or hypotheses and descriptions of the population. The time point of data collection is often irrelevant for
theory testing. If the concepts are adequately measured, the empirical analysis can also be based on older
data sets. It was particularly the economist Günther Schmölders, the other founding father of the ZA, who
has emphasized this economic advantage of a data archive.

Protecting Against Data Fabrication

A data archive is a protection against the fabrication of data. The problem of fake data is as old as
empirical research itself. If the reality does not fit the theory, make it fit by faking the data. The incentive to
fake may be particularly large in experimental empirical research because the data sets are of manageable
size. However, it also exists in the non-experimental sciences. And there are shades of grey! Is it already a
fabrication of data if a researcher attempts to increase the number of respondents in underrepresented
subpopulations? You have 45 percent men in the sample but you should have 50 percent. Therefore you
increase the number of men by simply doubling the interview records of some male respondents. As long
as the scientist gives a detailed report of what he did, you may call it a crude form of weighting even
though it can make the analysis of sample biases very difficult. If this “weighting” is tacitly done, however,
it becomes a fake. A data archive is certainly not a perfect protection against fake data, but it increases the
risk of detection and therefore the expected costs. If a data set is accessible to the scientific community,
everyone can examine its quality. This can be done with more or less sophisticated methods. Faking is not
as easy as it might appear! I assume that open access to the data will deter quite a few potential fakers. A
commission of the German Research Foundation (GRF) which has investigated means for improving the
quality of data has therefore recommended a binding obligation for data archiving: The data of a project
which is funded by the GRF have to be delivered to the ZA at the end of the project.

Professionalizing Empirical Research

The above considerations already indicate that the institutionalization of a data archive can and will
lead to the professionalization of empirical research. Even with good training in methods and statistics, a
young researcher will not have all the practical knowledge for carrying out a survey. What kind of
sampling procedure is the best given the funding of the project? How should the demographic questions
be phrased? Are there already well-established measurement instruments in Germany for the theoretical
concepts he is using? In these instances, he may ask a member of GESIS for advice. Accordingly, a data
archive contributes to the professionalization of empirical research by advising new research. This can be
done in very different forms: The archive can advise researchers by telephone or on a project site, the
archive can also actively participate in preparing and designing a study, and the archive can co-operate
like any other scientific institution in a project.

Supporting Reliability and Validity Analyses

As we all know, a good measurement instrument should measure what it intends to measure; it should
be, as we say, valid. Furthermore, the property should be measured in such a way that different observers
under similar conditions arrive at the same result. Those who are experienced in social research know that
the quality of a good measurement instrument can never be demonstrated on the basis of a single data set.
The danger of capitalizing on chance is simply too large. We have to test the instrument repeatedly in
order to harden it. And what may appear as a very promising operationalization after the first analysis
sometimes does not survive the next replication. Ambitious reliability analyses are often carried out by
several researchers who do not co-operate with each other. They order the data, evaluate the existing
literature and carry out their own analyses. The final measurement instrument then is the result of a long
series of analyses and tests. A data archive is almost indispensable for these advanced reliability and validity
analyses. And it is useful if reliable scales are later added to an existing survey.

Harmonizing and Standardizing Data

Closely related to the improvement of measurement is the harmonization and standardization of data.
Social scientists have a tendency to create their own measurement instruments and scales. They rarely are satisfied with the conceptualizations and empirical interpretations of their colleagues. This is certainly one of the core problems of qualitative research, however, it also exists in quantitative studies. To enter new questions into a questionnaire indicates creativity and imagination. As a consequence, the results of different surveys are hardly comparable. Suppose that an attitude survey was conducted in 1990 and another one 10 years later. Both surveys try to measure the same attitude, however, with slightly different items. In such a case we can never be sure that a percentage difference between the two attitude items really reflects attitude change. It can be due to the difference in item wording as well. To be sure, the same wording of the question is not a sufficient condition for demonstrating attitude change because an attitude question may also change its meaning. The question is simply understood in a different way than 10 years before. The some phrasing of the question is, in other words a necessary condition for the empirical measurement of value and attitude change but it is not sufficient. The ZA has continuously appealed to the social scientists in Germany to carefully examine the set of existing items and scales before they create new ones. Our efforts have so far had, as I have to admit, only limited success. The idea of replication is probably realized best in research programs like the General Social Survey ALLBUS, a joint enterprise of ZUMA and ZA, where the same questions are repeated in successive surveys. There is some hope that we will be able to establish a similar research program for the German election studies. However there is still much to be done.

The problem of standardization and harmonization is particularly relevant for international surveys. By standardization we hope to define equivalent measures for the same concept in different cultures. If we ask respondents about their age, we can hope to measure everywhere the same property, i.e. the biological age. If we measure marital status, things already become a little bit more subtle. Is the meaning of “being separated”, for example, really always the same? In some cultures, separation may be negatively stigmatized to such an extent that it will always be heavily underreported in surveys. It will, so to speak, exist as a social phenomenon but it will not exist in the surveys. Obviously, this is an instance where qualitative research becomes important. It can help us to understand why we find a difference between the actual and the reported behavior. However, the knowledge which we get from qualitative research is inevitably incomplete. Having analyzed 5, 10 or perhaps twenty cases, the qualitative researcher can never tell us whether he has analyzed a socially significant phenomenon or not. Social significance is also dependent, but not exclusively, on the frequency of the phenomenon, i.e. on numbers. This is, in my view, a strong reason, why quantitative and qualitative researchers have to co-operate. The results of qualitative research may enable us to develop better measurement instruments for quantitative research.

Marital status is only one of many illustrations for the problem of harmonization and standardization. There are other and also more complicated ones, social status for example. In order to measure social status in its “occidental meaning” you need occupation, occupational prestige, and income. Some of my Japanese colleagues tell me that you cannot ask income questions in Japan. If we take this for granted and believe at the same time that social status is an important concept for understanding and explaining Japanese society, we are immediately confronted with a number of complicated theoretical and methodological issues. If we cannot use income, can we use other, functionally equivalent measures instead? And what is the criterion for functional equivalence? Is it only external validity or are there other criteria which have to be met? Let me finally mention education. All economically advanced societies have developed highly differentiated systems of education. There is not only primary, secondary, or tertiary education, there is also continuous education, vocational training, etc. Regarding these complexities, the question arises as to what we intend to measure by education. Is it a surrogate for intelligence, is it knowledge, is it the socialization of values and beliefs, or is it something else? If education is equated to socialization during the formative years which may, let us be generous, last until 25, continuous education obviously is unimportant. If we are interested in the socialization of general moral norms and values, vocational training may be less relevant. By defining step by step the concept of education more precisely, we may finally arrive at a better comparable concept of education. This is certainly a task which cannot be
done by the ZA alone. At best, we may have a researcher in house who is interested in education. Then we may motivate him to engage in an international project on the measurement of education. If not, we can try to remove some of the inconsistencies from the education variables, and carefully document the differences in the education systems in our international surveys. And we can appeal to the scientific community to look for a solution of these measurement problems. By this, we can at least make people aware of the problems which still have to be solved. And sensitivity for the problems is the first step to higher professionalism.

Stimulating Co-operation

A data archive stimulates the co-operation between statisticians and empirical researchers. Statisticians often prefer to work with hypothetical data, if they work with data at all. They for a long time considered applied statistics inferior to pure statistics, at least in Germany. Apart from that, statisticians and social scientists do not speak the same language. There is a problem of communication. The ZA has made continuous efforts to bring these two groups together by organizing conferences and seminars and by trying to stimulate co-operation. In the long run we have made some progress even though the co-operation still can be improved.

Providing Education

The ZA does not only provide the scientific community with important information, it also has a very important education function. In the sixties, after the foundation of the ZA, Cologne was one of the leading centers, if not the leading center, in empirical research in Germany. The typical German social scientist knew neither factor analysis nor multiple regression analysis nor could he or she use statistical packages like SPSS. It therefore become an important task of the ZA to train young German social scientists in these methods, models, and programs. For this purpose, the spring seminar was established in 1972 which soon become a very successful program. During three weeks, young graduates, postgraduates, and scientists are trained in advanced statistical methods. They learn to apply these methods to empirical data—usually data from the ZA. In addition, these scholars are made familiar with the treasures of the data archive, its data sets, and its other facilities. Three weeks of hard work are sufficient to create permanent networks of young scholars who continue to keep contact with each other and often also co-operate with each other. The teachers in our spring seminars always have been international. A few have come from Germany, but many of them from Western Europe and abroad. The composition of the participants, however, has changed over time. While they were mainly from Germany in the beginning, nowadays about 50 percent of the participants are from other countries. All courses are given in English.

The spring seminar was an important first step to disseminating methodological and statistical knowledge in West Germany. It was later supplemented by an autumn seminar of our historical department which was mainly visited by students of social history and people with an interest in longitudinal analysis. The seminars have also become a model for courses in methods and statistics at German universities. The participants of our spring seminars understand that it is much more attractive to learn statistics and methods with real data. When they become teachers at their home universities they therefore begin to order data sets from the ZA and to use them in their courses. Nowadays, ALLBUS data or data from German election studies are used in courses on methods and statistics at many universities.

Planning for Continuous Education

So far I have only mentioned functions and activities which are actually performed by the ZA in Cologne. There are other tasks which are either in statu nascendi, as it is expressed in Latin, or have so far not materialized at all. A first is to add to our system of workshops, courses and spring seminars a program for continuous education. The ZA and GESIS have continuously co-operated with the Association of Social Science and Marketing Research Institutes. They organize a joint conference every year, and they also organize joint workshops and seminars. With the rapid growth of information technologies and
methodological knowledge, these activities could be bundled into a program of continuous education in empirical research.

While this program would address social scientists outside the universities, one may also think of courses for journalists. In Germany, the methodological and statistical knowledge of journalists is low. They are hardly familiar with the calculation of percentages and the majority of them have never heard about confidence intervals. As a result of this lack of knowledge, it is not surprising at all that they often give incomplete or even false information about empirical research. In order to change this, the ZA and ZUMA could offer regular courses for this clientele.

The training of journalists would be an important step toward the improvement of social knowledge in the society at large. A bigger step would include the preparation of public use files for schools. I am well aware of the problems which are involved. It would by no means be sufficient to provide schools with a data set. It would require special training for teachers and also some introductory textbooks for the pupils. However, if we do not begin very early to teach children also the elementary rules of empirical research, social sciences at school will remain in the hands of those who have never learned it. In some parts of Germany, the discipline is taught by historians or teachers of German language.

Concluding Remark

It is a big mistake to regard an archive as an institution for the distribution of data sets. In contrast to a library, an archive does not simply store and catalogue the data. Rather, it usually adds value by cleaning the data, by adding a study description and a methodological report, and by including indices, scales, and other measurement instruments in the data set. In order to perform these tasks an archive needs a highly competent scientific staff. Only if the archive can compete with other universities in terms of research quality and publications, will it be acknowledged in the scientific community. The higher the scientific reputation of an archive, the more attractive it becomes for the researchers to deliver their data to the archive, the easier it is for the archives to stimulate co-operation among scientists and to set a process into motion which finally leads to a culture of data sharing.