Measurement and selection bias in longitudinal data: a framework for re-opening the discussion on data quality and generalizability of social bookkeeping data

Baur, Nina

Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data

Nina Baur [*]

Abstract: *»Datenqualität und Verallgemeinerbarkeit. Eine Rahmen für die Wiedereröffnung der Debatte über Messfehler und Stichprobenprobleme von Massendaten«*. The author compares mass data with survey data and other process-generated data and discusses their relevance for historical, historical social science and sociological research. After summarizing the current state of methodological knowledge on public administrational data, she concludes that the discussion on mass data has to be re-opened. She suggests a framework for such a discussion and links the older German discussion from the 1970s and 1980s to the discussion newly arising. She suggests that the major issues are (a) data lore and measurement quality; (b) data selection and sampling problems; (c) archiving and statistical programmes and (d) data preparation. After summing up the state of the debate, the authors suggests which questions should be answered in future research.

Keywords: Longitudinal Analysis, Process-Generated Data, Process-Produced Data, Social Bookkeeping Data, Public Administrational Data, Mass Data, Mass Files, Survey Data, Data Lore, Measurement, Data Quality, Data Selection Bias, Sampling, Archiving, Accessing Data, Data Preparation.

1. Introduction

As the discussion in the HSR Special Issue on "Theory and Data" has shown (Baur 2009a), there is no one perfect data type. Instead, many factors contribute to which data type is best suited for social science research, e.g. the researcher's theoretical framework, the nature of the research question and the availability and quality of data. However, these factors themselves are subject to social change: In different historical times, different research questions are deemed relevant, different theories are preferred by the academic community,

[*] Address all communications to: Nina Baur, Technische Universität Berlin, Institut für Soziologie, Fachgebiet Methoden der empirischen Sozialforschung, Franklinstr. 28/29 (Sekr. FR 2-5), 10587 Berlin, Germany; e-mail: nina.baur@tu-berlin.de.
My special thanks to Wilhelm Heinz Schröder for helping me to reconstruct the debate on process-produced data of the 1970s and 1980s.

and different data are produced and available. This in turn means that the question of which data type is best suited for social science research has continuously to be asked anew (Scheuch 1977).

1.1 On the Nature of Mass Data

One of the oldest data types used in social science research are data arising from social bookkeeping, i.e. files produced for example during modern governments' administrational processes. While traditional object-oriented files ("Sachakten") are organized in a way that they collect information on a certain topic or object, for *social bookkeeping data/ public administrational data*, the data body is divided into parts or cases, e.g. a person or an object (a house, a car etc.) (Bick/Müller 1984: 123-124). Data collection is organized that the same type of information is collected on many cases. Thus these *quantitative/ standardized process-produced/ process-generated data* were called *mass data* ("Massendaten") or *mass files* ("Massenakten") in the German discussion in the 1970s and 1980s.[1] Another aspect of mass data is that, before information on a case can be collected, the case has to be registered at the data producing agency, e.g. a person has to be registered as a client at a government agency. After registration, mass data are organized as case histories, i.e. data producers collect as much information (variables) as possible on single cases (Bick/ Müller 1984: 124-125). This means, that typically, mass data are longitudinal data. In a lot of ways, mass data therefore resemble survey panel-data.

1.2 On the Relevance of Mass Data Compared to Other Data Types

Whenever *historians and historical social researchers* have applied quantitative methods, they typically have used mass data. In contrast, in *sociology* the relevance of mass data (in comparison to other data types) as a source in empirical research has been shifting historically, culturally and between research fields. Two of the earliest and most famous examples of sociologists using social bookkeeping data are Émile Durkheim's (1897) suicide studies and Max Weber's (1906-1922) study on the rise of modern capitalism.

After discovering random sampling and survey methodology, there was a phase from the 1930s to the 1950s, when using process-produced data declined within social science research. Although survey research has been the dominant methodology ever since, starting in the 1970s, process-produced data received

[1] For an overview over the history of quantitative historical research in Germany in it's relation to sociology, see Best/Schröder 1987, Best 1996, 2008a, 2008b. Note also that there are different international styles of doing quantitative history (Clubb/Scheuch 1980, Jarausch 1987) and that this may well influence the way mass data are handled in research practice. For a systematic comparison on different strands of historical social research in the international debate, see Schröder (1988, 1994).

increasing attention, especially in countries like Germany which are historically rich in public administrational data. Some reasons for this resurgence of using social bookkeeping data were:

1) *Paradigm Shift within Methodology:* Some fields of research like criminology shifted from a qualitative to a quantitative paradigm (Bick/Müller 1984: 125-126).
2) *Paradigm Shift within Theory:* Within these research fields, research questions shifted (Bick/Müller 1984: 125-126). E.g. researchers have been increasingly interested in institutions and in longitudinal research. Especially for longitudinal questions, process-produced data often are the only option, as no research-elicited data exist for many research question – simply, because no-one thought the question important enough thirty years ago to collect data on it (Baur 2004, 2005).
3) *Development of IT:* There have been enormous advances in IT, facilitating data management and making preparation and analyses of large-scale administrational data possible for a broad range of researchers which would not even have been thinkable 40 years ago (Baur 2005).
4) *Accessibility:* Public administrations have started making data easily accessible for a broad range of researchers.[2] Examples in Germany are data made available by institutions such as the Federal Statistical Office, the Regional Statistical Offices,[3] the IAB[4] and the FDZ-RV.[5] In order to improve the development of availability of so-called micro data, even an own council, the RatSWD,[6] has been created, and workshops invite and introduce students and researchers to use these data, which are often already prepared for data analysis by the data-providing institution.

1.3 The State of the Methodological Discussion on Mass Data

While public administrational data share many problems with survey data, they also have some specific problems (Rokkan 1976). In contrast to research-elicited data (such as survey data), data production is not controlled by the researcher, but primarily for other reasons, e.g. for public administration purposes. An unknown number of factors possibly decreases data and sampling quality, e.g. individual strategies of clerks at the data producing agency in handling the data; the purpose the data were made for or the clients' strategies.

[2] An overview over data-producing institutions can be found in Baur/Fromm (2008a).

[3] "Statistisches Bundesamt" and "Statistische Landesämter", www.destatis.de.

[4] Institute of Employment Research ("Institut für Arbeitsmarkt- und Berufsforschung"), www.iab.de.

[5] Research Centre of the German Pension Fund ("Forschungsdatenzentrum der Deutschen Rentenversicherung"), http://forschung.deutsche-rentenversicherung.de.

[6] German Council for Social and Economic Data ("Rat für Sozial- und Wirtschaftsdaten"), www.ratswd.de.

Thus, the sampling and data collection process is not methodologically conceptualised, conducted and controlled by a researcher (Bick/Müller 1984: 124-125). As a result, data production is interwoven with society and institutions and subject to historical change. Societal and institutional filter influence

1) which data are produced and how they are produced (*production bias*),
2) if and how data are stored (*selection bias*).

This in turn means, that methodological problems of social bookkeeping data might wholly or partly differ from those of survey data. After scientific re-analysis of mass data re-surged in the 1970s, these special problems became evident. This in turn induced an intense methodological discussion on data quality and data management of public administrational data from the mid-1970s to the mid-1980s in German historical social research, namely in HSR, and many of the papers written then are still valid and worth reading today.[7] Three major issues were (1) factors biasing data production, (2) computing and software and (3) how to build and organize digital archives. Some protagonists in the discussion were Heinrich Best, Wolfgang Bick, Reinhard Mann, Paul J. Müller, Erwin K. Scheuch, Wilhelm H. Schröder and Manfred Thaller.

After this first intense methodological debate, discussion has dwindled and methodological problems have only been rarely discussed within *historical social researc*h. In *sociology*, methodological discussions have focused more strongly on data analysis (particularly statistical analysis procedures) and on research-elicited data (particularly survey data).[8] This has lead to the paradox situation, that while probably more researchers are actually using mass data in empirical research than ever before, surprisingly little methodological knowledge exists on how to properly handle them during the research process:

Introductions to social research (e.g. Behnke et al. 2006, Diekmann 2006, Kromrey 2006, Schnell et al. 2005) typically strongly focus on how to accurately collect and use research-elicited data (Baur 2005). Mass data are usually summarized with other process-produced data in the category “documentary analysis” (“Dokumentenanalyse”) or content analysis (“Inhaltsanalyse”). Discussions on these data are either very short or missing completely at all. In contrast to discussions on survey methodology, they are definitively too short to enable students to handle these data in research (Ludwig-Mayerhofer 2003).

While traditional social science research either propagates qualitative research or quantitative research, the newly arising *Mixed Methods Research*

[7] I have compiled a list of these older articles still worth reading in HSR Trans 22 http://hsr-trans.zhsf.uni-koeln.de. The most important articles of the discussion are reprinted in Schröder 2006.

[8] Survey methodology is now a whole research field, dedicated solely to establishing factors decreasing sample and data quality and to develop procedures to either avoid these factors or – if this is not possible – to handle them professionally. Some leading centres of survey methodology are the Universität Mannheim (Germany), the University of Nebraska-Lincoln (USA) and the University of Michigan (USA).

(*MMR*) bridges the differences between these research traditions. Currently, MMR mainly focuses on problems of research designs and sampling strategies suitable for mixed methods designs (Teddlie/Tashakkori 1998, Creswell/Plano Clark 2006, Creswell et al. 2008, Journal of Mixed Methods Research) – discussions on data quality are rare. Although process-generated data are generally stated as one possible data type for mixed methods designs (e.g. Johnson/Turner 2003, Hunter/Brewer 2003), methodological debates usually only discuss how to mix research-elicited qualitative and quantitative data. In contrast, process-produced data's specific stengths and weaknesses are typically neither discussed, nor are guidelines given how to handle them in research practice (Baur 2005: 87; 280-281, 319). The little information that exists on process-generated data is not much help for research practice.

This gap between knowledge on research-elicited and process-produced data is not only reflected in introductions to social science methodology, but also in *advanced methodological discussions*. For example, Sage has started publishing key articles on specific methodological topics at the turn of the century. Each issue of the "Sage Benchmarks in Social Research Methods" consists of four volumes. Articles selected for the volumes are supposed to reflect the state of the art of the Anglo-Saxon discussion on social science methodology. Eight issues of the series directly focus on survey methodology (De Vaus 2002, 2007, Bulmer 2004, 2010, Scott/Xie 2005, Bartholomew 2006, Roberts 2008, Bulmer et al. 2009), reflecting the depth of knowledge we today have in this field.

In contrast, no single volume specifically focuses on social bookkeeping data. Papers on mass data are sprawled either in the single issues on "Historical Methods in the Social Sciences" (Hall/Bryant 2005) or "Documentary Research" (Scott 2006). When looking at the volume on "*Documentary Research*" (Scott 2006), it becomes clear, that in the Anglo-Saxon tradition, process-produced data and documentary research are associated mainly with qualitative research. Major issues of discussion are life and personal documents (e.g. autobiographies, diaries, and letters), visual materials (e.g. photographs and videos) and archaeological artefacts. The little literature existing on mass data focuses on censuses and official statistics. The only article included in the four-volume set explicitly discussing data quality of mass data is Platt (1981).

In contrast, books on *historical sociology and historical social research* typically introduce major theories (Spohn 2000, Gelanty/Isin 2003, Schützeichel 2004, for the German discussion see Baur 2005). Discussions on methodology typically remain epistemological or ontological. Authors argue why one should use historical methods at all; if to use a interpretative or positivist paradigm; how to build theory; how to handle the problem of time etc. (e.g. Ruloff 1985, Best 1988, Tuchman 1994, Hall 2007). In comparison to discussions on survey methodology, these treatises of methods of social research remain on a relatively general level, which is important and interesting to read but gives no or little guidelines of how to use data in actual research.

1.4 Why a Renewal of the Discussion is Needed

This lack of methodological debates on social bookkeeping data would be excusable, if there were not a number of factors that stress the need for such a discussion:

1) *Danger of Knowledge Loss Due to Generational Change:* The initiators and protagonists of the debate of the 1970s and 1980s have either already retired or will be retiring within the next years. This type of generational change always carries the danger of a loss of existing knowledge. In recent years, this danger has even increased, as (a) due the logic of the international academic publication market researchers increasingly read only newer literature and tend to neglect older literature, and (b) researchers increasingly rely on electronic data bases for finding literature. However, a lot of the older literature cannot be found with these data bases. (c) Additionally, the research fields often applying mass data have shifted from e.g. criminology, sociology of deviance, sociology of the law and urban sociology in the 1970s and 1980s to life course analysis, social policy and labour market research today. As most researchers primarily read literature in their own field of research, the danger of losing existing knowledge is multiplied.
2) *Open Questions:* The debate of the 1970s and 1980s sensitized researchers for some problems, but the discussion ended before these problems were solved and before results were systematically integrated.
3) *New Research Fields:* The early debate focussed on data suitable for fields of research then prominent, namely demography, public administrations, criminology and deviance, cities and regions. Today, research has shifted to other fields that were not deemed interesting then by many researchers, e.g. research on the media, organizations, consumer, producer and labour markets. It is yet unclear, which types of process-produced data are suitable for these newer fields, how these data are distorted and if the rules for handling distortions developed for data in the earlier research fields apply for data in the newer research fields.
4) *Technological Change:* As many of the discussions of the 1970s and 1980s were on technological issues, results may have become obsolete due to technological change. Instead, new technological problems might have arisen. At the same time, technological change has created a new problem, as it has facilitated both analysis of process-produced data and secondary analyses of survey data. Thus, more and more researchers who are very often less experienced than the earlier researchers analyze data which they have not collected themselves.
5) *Disassociation of Data Collection and Data Analysis:* This disassociation between data collection and data analysis can result in researchers presuming that they are working with “hard”, “neutral” facts, as they are not aware of or do not acknowledge distortions of the data (Waldow 2001).

6) *Easier Access to Mass Data:* As stated above, public administrations increasingly facilitate access to their data, which has given research using these data a new impulse. However, this easy access also carries the danger that researchers use these data without being aware how they are biased and restricted.
7) *Internationalization:* Social researchers increasingly conduct international comparison. Very often, they know only little of some of the countries of comparison, e.g. not even the primary language. In these cases, mass and survey data and results are de-contextualized from the original narratives, i.e. the process of data collection. Again, this carries the danger of researchers working with distorted data without being aware of the problem (Waldow 2001, Baur 2009c).
8) *Increase of Longitudinal Research:* In a way, sociology has moved towards historical sociology and historical social research, as many of the newer research questions can only be answered when doing longitudinal research, e.g. in life course, educational, social policy and labour market research. Regardless, if researchers are using process-produced data or are reanalyzing research-elicited data, longitudinal research poses some special problems, e.g. what happens if the boundary of the case or the population changes (Abbott 2001, Baur 2004). Although most of the actual research conducted by social historians has always been longitudinal the methodological debate in the 1970s and 1980s never came around tackling these problems, as it got stuck in earlier stages of the research process.

These are some of the reasons, why with this special issue we want to reopen the discussion on data quality and generalizability of public administrational data. Our aims differ from the first wave of discussion on mass data in an important aspect: The discussion of the 1970s and 1980s was driven by a *source-perspective*: Authors described and characterized specific source types, e.g. criminal statistics, labour-market data, census data etc. (Müller 1977). Bick et al. (1984a) even demanded a sociological source lore ("Soziologische Quellenkunde") or data lore ("Datenkunde"). Starting from these data, authors asked about specific sampling and measurement problems. Today, we aim at taking a *research process-perspective*: Starting from different stages of the research process, we ask, what can happen there with and to data and samples? Which problems can arise? Can they be avoided? If not, how should they be properly handled? Do these problems arise for all sources, or are they specific to some sources? The advantages of such a perspective are:

1) It is easier to transfer knowledge from one research field to another and to compare results across sources.
2) The debate can be linked to other methodological debates more easily, e.g. those on survey methodology and on mixed methods research.

In order to structure the discussion, in chapter 2 of this paper I will thus give an overview over the research process of standardized process-produced data (mass data) and compare it with standardized research-elicited data (survey data). Using this model of the research process, I will show that there are four main points of the debate. In chapters 3 to 6, I will discuss each of these points and try to link the older debate with the newer debate in this special issue. As can be seen from the papers in this special issue, we are still at the very beginning of this second wave of discussion: We still have to find the right writing style for a research process-oriented discussion. We are still grappling at saving the older knowledge and at defining the new questions. We ask more questions than can be answered, and many of the answers found are preliminary, as they have to be tested with other data in other countries. I thus will conclude with open questions for future research.

2. The Research Process for Quantitative Longitudinal Analyses

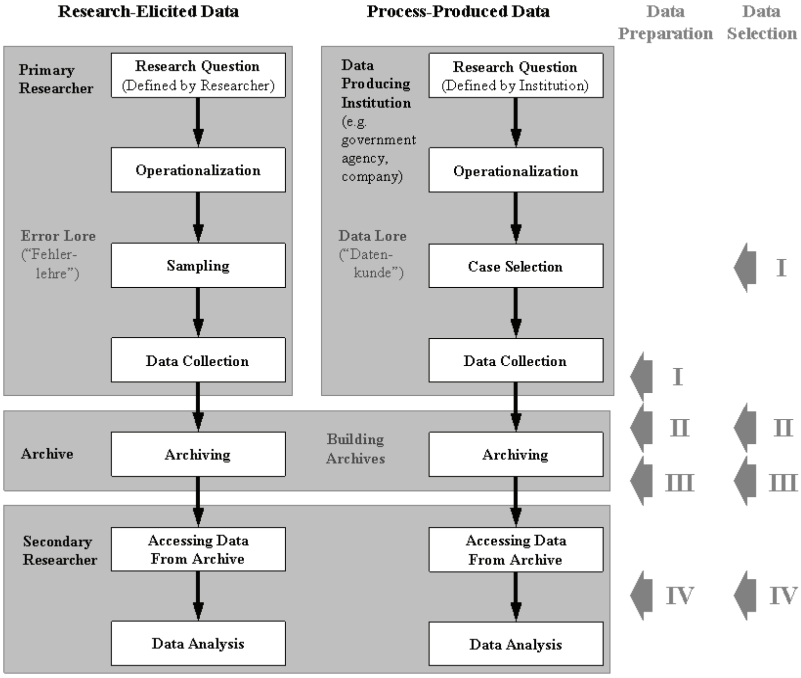
The type of research-elicited data most similar to process-produced mass data are survey data. As survey data are also the data type most intensively discussed in methodological literature, I suggest structuring the debate on mass data by comparing them to survey data and asking which types of problems researchers typically have to address during the research process. I will thus start with comparing the research process for survey data with that of process-generated data. Graph 1 gives an overview over the steps to be taken for both research-elicited and process-produced standardized data.

2.1 Research-Elicited Standardized Data (Survey Data)

In surveys, the primary the researcher start from a research question and theoretical concepts. Next, he develops a research design, operationalizes theoretical concepts into survey questions and designs a questionnaire (Hox 1997). At the same time, the researcher has to define a survey population and draw a sample. Ideally, this should be a random sample, as random samples enable the researcher to use inferential statistics later on and compute probability of errors.[9] Theory plays an important role in this process, as it both defines which kind of data are best suited for analysis and how concepts can and should be operationalized (Baur 2009b).

[9] For an overview of the process of survey design and of typical errors, see: King et al. 1994, De Vaus 2002, Bulmer 2003, Fink 2003; Groves et al. 2004; Behnke et al. 2006, Diekmann 2006, Kromrey 2006.

Graph 1: Typical Problems to be Solved During the Research Process in Longitudinal Research



The next steps after operationalization and instrument design are data collection and data preparation. Data preparation covers transforming variables, fusing data (= linking record) and deleting cases. Data preparation can take place during at least four steps of the research process:

1) *Data Preparation I: Preparing Data During Data Collection.* Data some times are prepared concurrently or immediately after data collection by the primary researcher. E.g., if an interviewer has been caught forging interviews, the primary researcher will typically delete or at these mark all interviews conducted by the culprit immediately.
2) *Data Preparation II: Preparing Data for Archiving.* Longitudinal data *have* to be archived, as data collection takes place over a long period of time. In many countries, cross-sectional surveys are also increasingly archived. Usually, before data are archived, they are prepared for the archive's special needs. Examples for such transformations are renaming variables; deleting cases or variables that are deemed unnecessary for archiving; deleting false information, transforming data into a common format, deleting information that for some reason cannot be archived (e.g. due to structure of the archive,

due to archiving space). Archiving can also mean that information is added, e.g. documentation of the data generation process.

3) *Data Preparation III: Preparing Data for Release From the Archive*. Before leaving an archive for secondary analysis, data are usually prepared again. Some of the reasons for data transformations are: deleting inconsistencies; deleting cases or variables the secondary researcher does not need or should for some reasons not process, e.g. due to data protection laws ("Datenschutz"); preparing data in order to make it easier for the secondary researcher to use them. Very often, the latter means that only a sub-sample leaved the archive.
4) *Data Preparation IV: Preparing Data During Data Analysis*. Not only the archiving institution, also the secondary researcher often transforms data for her specific aims of data analysis. Typical examples are recoding variables, computing new variables and further reducing the sample to the group interesting for analysis.

Only now the researcher can start with data analysis. Because the survey research process is organized as a linear process with one step being completed before the next one begins and because data often are not collected by the same person who will analyse them later, there is a danger of de-contextualising data analysis – not only from the socio-historical context in which the data were collected (Baur 2009c) but also from the process of data generation (Waldow 2001). In fact, both data collection and sampling are closely connected to data analysis:

1) *Relation between Questionnaire and Descriptive Statistics:* Certain question formats and survey design imply certain descriptive statistical analysis procedures and inhibit others, and these procedures aim at different theoretical goals (Baur 2005, Baur/Lamnek 2007). For example, factor analysis can only be conducted, if a list of theoretically and thematically connected items has been included in the questionnaire.
2) *Relation between Random Sample and Inferential Statistics:* A prerequisite for applying inferential statistics (testing hypotheses and estimating confidence intervals) is *always* a random sample. Inferential statistics are such a powerful tool, as the enable the researcher to estimate how likely it is that he has drawn false conclusions from the data (Biemer/Lyberg 2003: 305-350; Groves 2004: 239-294). Any drop-outs and missing information – regardless if they arise from people not being included in the sampling frame, from unit nonresponse or from item nonresponse – may bias the results.

While questionnaire design is something the researcher can control (almost) completely, sampling quality is increasingly becoming a problem. Generally, data selection (other than random sampling) can take place during at least four stages of the research process:

1) *Data Selection I: Data Production Bias*. General problems when interviewing people are that the researcher (a) has to include people in the sampling

frame, (b) has be able to establish contact with them, and (c) they have to be willing to participate in the survey and be able and willing to give correct information. Decreasing response rates demonstrate that this is becoming less and less the case. Unfortunately, nonresponse is usually not random. Instead, only certain types of people drop out, and the reason they drop out for is almost always related to the research question.

2) *Data Selection II: Selection of Cases for Archiving*. In order for data to be archived, an archive has to exist and the archiving institution has to decide that the dataset is worth archiving. It still might decide not to archive specific cases or simply make mistakes during archiving. All these processes may further bias available data for future research.
3) *Data Selection III: Making Data Accessible for Researchers*. Data can be further selected, as secondary researchers might not gain access to data: They might not know the archive exists, they might not know about the existence of a specific data set, and they might not get access to the dataset or to specific variables. For example, in Germany GESIS[10] archives many social science data. However, some data are only available to academic research (not for commercial research), and researchers do only get direct access to sensitive data, if they personally travel to GESIS, have been approved for data access in a formal process and do analyse the data there.
4) *Data Selection IV: Selection of Cases for Analysis*. Researchers might purposefully delete cases from the data set for data analysis, e.g. they might concentrate on one specific sub-group or delete outliers.

Note that almost none of these procedures are faulty in themselves. Most of them are legitimate steps taken during the research for a reason. However, (a) mistakes may happen during this process and (b) the more data are tailored for one purpose of analysis, the more useless they might become for other purposes, i.e. they might be biased too much for the latter purposes. Survey research thus has discussed systematically when and how errors may arise during the survey process and developed an *Error Lore* ("*Fehlerkunde*"). Groves (2004: vi) summarizes that "sample surveys are subject to various types of errors:

1) *Coverage error*, from failure to give any chance of sample selection to some persons in the population.
2) *Nonresponse error*, from failure to collect data on all persons in the sample.
3) *Sampling error*, from heterogeneity in the survey measures among persons in the population.
4) *Measurement error*, from inaccuracies in responses recorded on the survey instruments. These arise from:

- a) effects of *interviewers* in the respondents' answers to survey questions;

[10] "GESIS – Leibniz Institute for the Social Sciences", www.gesis.org.

- b) error due to *respondents*, from inability to answer questions, lack of requisite effort to obtain the correct answer, or other psychological factors;
- c) error due to the weaknesses in the wording of survey *questionnaires*; and
- d) error due to the effects of the *mode of data collection*, the use of face to face or telephone communication" (Groves 2004: vi).

Survey methodology discusses how to assess, when and if survey errors occur; how to estimate the size and effect of the error; how to avoid these errors and – if avoiding them is not possible – how to handle them. The extensive body of literature that has been produced within the last decades covers topics like coverage error (Biemer/Lyberg 2003: 63-79; Groves 2004: 81-132), nonresponse error (Brehm 1993, ADM 2004, Koch/Porst 1998, Schnell 2000, Groves et al. 2001; Biemer/Lyberg 2003: 80-115; Groves 2004: 133-238), measurement problems in general (Bartholomew 2006; Biemer/Lyberg 2003: 116-148, Biemer et al. 2004, Campbell/Russo 2001, Salkind 2006, Rost 2004) and how to measure specific concepts like age, ethnicity or occupation (Hoffmeyer-Zlotnik/Wolf 2003; Hoffmeyer-Zlotnik/Harkness 2005). Other issues discussed are how to correctly design a questionnaire; how the interviewer (Biemer/Lyberg 2003: 149-187 Groves 2004: 357-406), respondent (Groves 2004: 407-448) and data collection mode (Fuchs 1994; Dillman 2007; Biemer/Lyberg 2003: 188-214; Groves 2004: 501-552) may influence the data; and what mistakes may happen during data processing (Biemer/Lyberg 2003: 215-257). In other words, today we know very well how surveys do (not) work, and survey research has also established procedures to evaluate the survey process, ensure data quality and good practice (Prüfer 1996; Esposito/Rothgeb 1997; Presser et al. 2004).

2.2 Process-Produced Standardized Data (Mass Data)

As graph 1 on page 17 illustrates, for process-produced mass data, in principle the same steps have to be taken as for research-elicited standardized data. In order to operationalize a (1) research question, (2) a research design, standardized questions (similar to survey questions) and a form (similar to a questionnaire) have to be developed, (3) cases have to be selected for which data have to be (4) collected and (5) archived. Secondary researchers have to (6) gain access to the data and (7) analyze them. Equivalently to survey research, measurement quality and sampling are two major problems. Problems with data selection and data preparation occur at the same stages of the research process as in survey research, and for pretty much the same reasons.

The main differences between survey research and process-produced mass data is that in survey research, steps 1 to 4 are conducted by a researcher for scientific purposes, while for mass data, these steps are conducted by a data collecting institution (e.g. the government, a company) for purposes other than research, i.e. the data are tailored for that other practical purpose (Baur 2004).

Mass data thus have several advantages in comparison to survey data (Baur 2009b): They are non-reactive. They can be used, if other means of data collection are not applicable, for example, if infrastructure for large-scale surveys does not exist, if response-rates in surveys are expected to be to low, if researchers might not get access to interview partners or if the social phenomenon of interest is not observable (e.g. when analysing past events or hidden populations).

At the same time, mass data face several problems. The severest is that, while survey data *may*, mass data almost certainly *do* suffer from measurement problems and sampling bias. *How* they are they biased depends on (1) the particular purpose, (2) format and (3) institutional embeddedness of the pertinent data type. In addition, all three elements may change over time (Baur/Lahusen 2005). Finally, the original data might decay or will be destroyed deliberately. The process of data selection is biased, too, as humans have to actively want to preserve data available for later use (Baur 2004, Baur/Lahusen 2005). In other words, in order to assess how (much) data are distorted from reality, researchers exactly have to know how and why mass data were created (Bick/Müller 1984: 128) and how and why they were preserved. It is therefore no surprise that the discussion of the 1970s and 1980s focussed on the following topics:

1) *Data Lore and Measurement Quality:* What factors distort data and samples during data production? How do they distort data and how large is the effect? Because some of these processes are highly specific to a certain topic, country or data type, Wolfgang Bick and Paul J. Müller suggested establishing a sociological data lore ("Datenkunde") as equivalent to the error lore in survey research.
2) *Data Selection:* When and how is information selected? Should researchers sample data further for analysis, and how should they sample data? When and how can information drawn from mass data be generalized?
3) Archiving and Statistical Programmes: How should archives for mass data be built that information loss and distortion is minimized and that they are practical both for data suppliers and researchers? Which software and data base structures are best suited for preparing and analyzing mass data. This discussion was closely associated with the name of Manfred Thaller.
4) Data Preparation: As errors cannot be avoided, how can and should they be handled during data preparation?

The discussion on these topics was interwoven, as problems are related. In the next four chapters, I will try to summarize and assess the results of the earlier discussion and link it to the papers in this special issue.

3. Data Lore and Measurement Quality

As process-generated data are not generated for scientific but for practical purposes, in contrast to survey data, the question with process-produced data is not *if* they are biased but *how* they are biased during data production (Baur 2004). *How* they are they biased depends on the particular purpose, format and institutional embeddedness of the pertinent data type. In addition, all three elements may change over time (Baur/Lahusen 2005). In other words, process-produced data usually suffer from severe measurement problems.

Assessing and handling this process of distortion during data production thus was the major topic of the debate until the mid-1980s. The discussion was driven by a *source-perspective*: Authors described and characterized specific source types, e.g. criminal statistics, labour-market data, census data etc. (Müller 1977). Since only few articles have been added since then and public administrations have been collecting data sometimes for centuries with only minimal changes in administrational procedures, much of what has been said then about bias in mass data is still true today. At least for Germany, much of what can be said about the data types discussed in the 1970s has been already said then. Therefore, many of the articles are still worth reading today. I have compiled a list of these articles in HSR Trans 22 (http://hsr-trans.zhsf.uni-koeln.de). Table 1 gives an overview on which topics references can be found in the compilation.

How much sense such a discussion would also make for research-elicited data is shown by *Christoph Thonfeld* in this special issue. Thonfeld uses oral history data as an example and shows that before secondary analysis of data, researchers need to ask similar question as for process-generated data.

However, discussing the particular advantages and disadvantages of specific data was not what the discussion aimed at. Rather, Bick et al. (1984a) pointed out that similar to the "error lore" ("Fehlerkunde") in survey data, there should be a *sociological source lore* ("*Soziologische Quellenkunde*") or *data lore* ("*Datenkunde*"). Equivalent to research on survey methodology, there should be research on distortions of process-produced data during the data production process, the two major methodological problems being sampling errors (i.e. cases not being registered at administrations) and measurement problems (i.e. reality being falsely represented in data). The idea was (1) to develop a general frame in order to assess the specific characteristics of certain data types, and (2) to systematize the errors that could arise. The leading figures in this process of systematization were Wolfgang Bick and Paul J. Müller who pointed out that data could be distorted in at least three stages of the data production process (Bick/Müller 1984: 128):

Table 1: Selected Articles on Data Lore for Specific Data Types

Topic	Section in HSR Trans 22
General Methodological Discussions By Research Field	Section 1
Family and Genealogical Research	Section 2.1
Educational Research	Section 2.2
Occupations, Professions and (Elite) Careers	Section 2.3
Labour Market Research	Section 2.4
Social Policy Research	Section 2.5
Demography, Censuses & Official Statistics	Section 2.6
Geriatrics and Gerontology	Section 2.7
Medical Sociology	Section 2.8
Public Administrations	Section 2.9
Organizations, Economies and Markets	Section 2.10
Churches and Religion	Section 2.11
Politics and Demography	Section 2.12
Media	Section 2.13
Military	Section 2.14
National Socialism	Section 2.15
Criminology, Research on Deviance and the Legal System	Section 2.16
Localities, Cities and Regions	Section 2.17
By Characteristics of Data-Producing Institution	
Communal Data	Section 3.1
Regional Data	Section 3.2
Federal Data	Section 2.6

1) *First Contact (Selection Bias I):* Before the client first appears at the agency and data can be collected and stored, a lot of things may happen either in her personal environment or in course of the interaction between client and the agency that may prevent or bias data collection. For example, the client may not want to disclose (certain) information. The administrational rules or a single clerk may not consider the client as a case worth collecting data on. Several organizations or clerks may interact in a way that data production is prevented or distorted.
2) *Difference between Reality and Data (Measurement Errors):* Even if data are collected, reality may be incorrectly represented in the data. Measurement errors may be caused e.g. by different perspectives of clients and the

agency's personnel on the problem or by clerks accidently entering false data. Different agencies may collect data differently, which may cause problems when aggregating data. During data aggregation, selection bias may also cause validity problems, leading to underreporting ("Dunkelziffer").

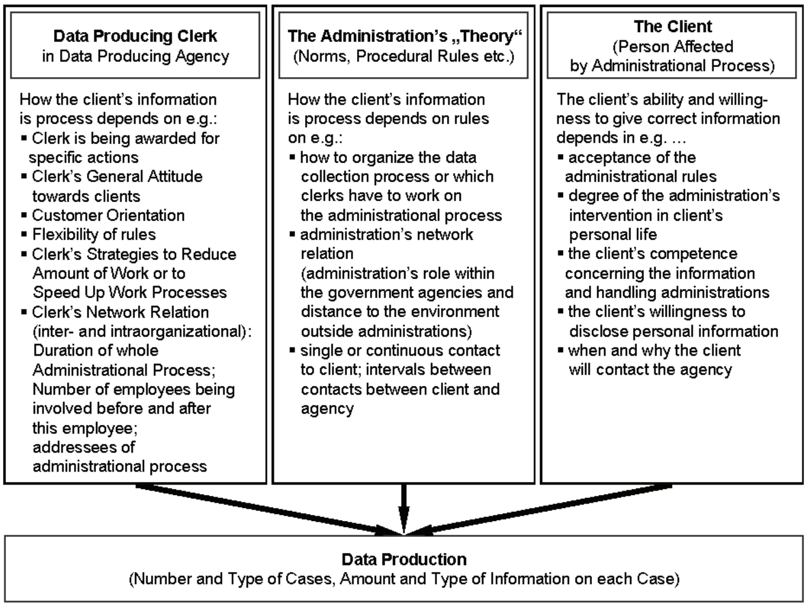
3) *Methodological and technical problems arising when using process-produced data (Selection Bias II):* Governmental files are often complex systems of data which have been collected at different points in time by different persons. Additionally, the have to be administered somehow – either electronically or by hand. When researchers want to use these data for their own research, a number of problems may arise from these factors, the first of them being sampling. Other problems are how to find data and get access to them, how to assess validity of data and how to fuse data with other data sets.

While discussing *when* data are distorted, another way of tackling the problem is to discuss *by whom* distortions are caused. Bick and Müller (1984) argued that there are three main sources of distortion: the Administration's "Theory of Reality", i.e. norms and formal procedures for data production; the data producing clerk and the client's behaviour. Graph 2 gives an overview over some ways of how these entities may distort data. Additionally, these factors may interact.

Many of these factors may affect each data type in a very specific way, making mass data fruitful for some research questions but not for others. It is yet unclear, if some general patterns can be determined that are valid across data types, institutions and nations. It also is an open question, how large the effects of certain types of distortions are and how to handle them in data analysis. Unfortunately, the discussion silted before results could be systematized.

Many of the authors in this special issue tie in with this earlier discussion on distortions of public administrational data during data production. In contrast to the debate of the 1970s and 1980s, they try to focus not on the data type, but on the factor causing the distortion; they try to estimate effect size and to develop solutions for handling these problems. The idea of tackling the problem this way is that this enables researchers to discuss these problems across data types. If this were done for more data types and countries, in the long run we could try to decide (a) which kinds of factors are really specific to a single data type and (b) which kinds of factors are predictable. I believe, that this would be exactly what Bick and Müller (1977, 1980, 1984) had in mind when they demanded a social science data lore.

Graph 2: Factors Possibly Contributing to Distortion of Data Production



Data Producing Clerk
in Data Producing Agency

How the client's information is process depends on e.g.:
- Clerk is being awarded for specific actions
- Clerk's General Attitude towards clients
- Customer Orientation
- Flexibility of rules
- Clerk's Strategies to Reduce Amount of Work or to Speed Up Work Processes
- Clerk's Network Relation (inter- and intraorganizational): Duration of whole Administrational Process; Number of employees being involved before and after this employee; addressees of administrational process

The Administration's „Theory"
(Norms, Procedural Rules etc.)

How the client's information is process depends on rules on e.g.:
- how to organize the data collection process or which clerks have to work on the administrational process
- administration's network relation (administration's role within the government agencies and distance to the environment outside administrations)
- single or continuous contact to client; intervals between contacts between client and agency

The Client
(Person Affected by Administrational Process)

The client's ability and willingness to give correct information depends in e.g. …
- acceptance of the administrational rules
- degree of the administration's intervention in client's personal life
- the client's competence concerning the information and handling administrations
- the client's willingness to disclose personal information
- when and why the client will contact the agency

Data Production
(Number and Type of Cases, Amount and Type of Information on each Case)

Source: Bick/Müller 1984a: 138, translated and adjusted by Nina Baur.

Most the authors in this special issue focus on the administrational "theory": *Tatjana Mika* discusses *how institutional filters work and how institutional changes affect data*. Applying the model introduced by Bick and Müller (1984), she develops a checklist of to judge the adequacy of process produced data for a particular research question. She argues that the same data can be adequate for some research question and inadequate for others. She illustrates her point by showing how two welfare reforms affect the numbers of caregivers and unemployed in the German pension fund data:

1) Concerning first time institutionalization, Mika's data show at certain historical moments an *increase/decrease in the number of registered cases* is caused by new institutional rules and regulations (and not by social change). This makes it difficult to compare age cohorts which were affected by social policy changes in different stages of their life course.
2) Due to eligibility rules, institutional filters also influence on which persons which type of information is gathered. For example, Mika's data show that due to changes in the unemployment assistance scheme, today *other types of persons now are registered* as unemployed than before the reforms.

Thomas Kruppe, *Peter B. Meyer* and *Gunnar Thorvaldsen* point out that administrations have to develop *concepts for measuring data*. How data are measured is usually defined by administrational procedures, and these procedures may and do change over time. Moreover, the definitions are not necessarily the ones a researcher would use. Regardless if the question is: “Who is unemployed in Germany?” (Kruppe), “Who has an occupation in the USA?” (Meyer) or “Which ethnic group does someone belong to in Norway?” (Thorvaldsen), the result is always the same: Whenever an administrational procedure changes, some people change from one category to another, although their life has not changed, i.e. people unreflectively using the data may assume that there was social change. Kruppe shows that these different definitions largely effect statistical results, i.e. which definition is used is by no means irrelevant. Meyer tries to assess which types of social groups are most affected by these definitional changes. Thorvaldsen tries to systematize these causes and suggest how to handle them.

Even if the definition of a social category remains the same, the *mode of collecting data* may change due to technological change. *Gunnar Thorvaldsen* discusses how changes in data collection mode are intertwined with other factors distorting census data. *Christian Seysen* shows that this aspect also has to be taken into account, as it also affects statistical results.

Bick and Müller (1984) also stated that different government agencies produce data for different ends and thus use different administrational rules for data production. They usually communicate badly or not at all. The result may be many inconsistencies between *data on the same person from various sources*. Using German employment histories as an example, *Markus Köhler* and *Ulrich Thomsen* show how these inconsistencies are caused and how they can be identified.

That *clients* may also influence the process of data production is long known, for example, data can be forged (Lippe 1998, Salheiser 2009), or people may not deem an aspect relevant to the problem and just forget to tell the data-producing agency. While it is possible to assess typical patterns of answering behaviour (as is illustrated by research on social desirability), these problems increase in longitudinal research, these patterns of answering questions or filling forms may change. It is known from survey research that social and cultural memory shapes the way respondents answer questions (Baur 2009c):

1) People answer depending on cultural knowledge and context, but this knowledge is not reflected in data.
2) The meaning of words can change over time or differ in cultures.
3) Hidden narratives form perception of question and answers.

Martina Huber and *Alexandra Schmucker* try to estimate the effect of the clients’ influence on data quality by triangulating survey and process-produced data on Germans’ employment careers. In a second step, they develop a typology of inconsistencies between both data types. In a third step, they try to assess for which types of clients wrong and inconsistent data are most likely. Concerning information on employment status, gender and nationality do not seem to have any effect. Until the age of 45, the likelihood of data being inconsistent decreases, afterward it increases again. Data are more likely to be correct for lowly educated persons and persons with a medium income. The most inconsistencies arise with people who were affected by many status changes in a short time. It is unclear, if the cause is that people with many status changed have a hard time remembering all status changes (then the mistake would be in the survey data) or if administrations have trouble keeping up with the changes (then the mistake would be in the administrational data).

4. Data Selection

Apart from measurement quality, sampling quality is a major problem both in survey and in mass data and therefore has been a major topic of discussion on social bookkeeping data since the 1970s. Although most survey samples are biased, survey researchers can control at least in theory the sample quality and – if they apply random sampling techniques – use inferential statistics later in order to compute the probability of drawing false conclusions from data. Mass data researchers do not share this luxury: mass data samples are almost certainly are *not* random samples. Instead, they are typically biased. Thus, it makes sense to discuss when and how cases are selected and how this influences sample quality. Generally, researchers have to take the following steps in order to assess if and how data can be generalized:

1) Defining the Population;
2) Assessing General Availability of Specific Data Formats and Searching Data;
3) Assessing and Handling Data Selection I: Data Production Bias;
4) Assessing and Handling Data Selection II: Selection of Cases for Archiving;
5) Assessing and Handling Data Selection III: Making Data Accessible for Researchers;
6) Data Selection IV: Selection of Cases for Analysis.

4.1 Defining the Population

The first step to be taken is defining the population for which the researchers wants to generalize results. This can quickly become more difficult as it seams because often, *populations do not have clear boundaries*. Additionally, in longitudinal research, *populations can change*. Most of the time, this change occurs subtly. For example, if the population are the inhabitants of a country, people can immigrate and emigrate, be born and die. At other times, changes of the population are sudden and drastic. A prominent example is German unification: Within a year, the geographical area called "Federal Republic of Germany" and the FRG's population increased dramatically, and since then, there has been a lot of within-country migration. At the same time, many administrational processes were adapted to the new situation. The methodological problem this poses is: (How) can data from before 1989 be compared with those after 1989? Does one analyze "Germany" as a whole, or does one analyze West and East Germany separately? If one wants to compare today's Germany, does one compare it to former East or to former West Germany (Baur 2004)?

4.2 General Availability of Specific Data Formats

Once the population is defined, researchers can start searching for data suitable for their research question and for this specific population. Researchers might be strongly limited in choice of data, as *depending on the historical period they want to do research on, only specific data might be available* (see table 2): It is a simple fact, that specific data formats were invented at different times:

If researchers today conduct interviews themselves, they are limited to the respondents' memory, i.e. (depending on the respondents' age) to about 70 years back from time of data collection Alternatively they can re-analyze survey data. However, this strongly limits the timeslot of analysis: panel-data have been collected in Germany only since the 1980s, trend-design data since the 1970s. The earliest cross-sectional data that could be replicated stem from the 1940s. In addition, these survey data only cover topics that were deemed interesting by researchers during the time of the first data collection. For example, for Germany many good survey samples on political behaviour exist, as after National Socialism, this was seen as an important topic. In contrast, educational behaviour has only been a topic of broad academic interest since the Pisa studies at the turn of the millennia. Thus, a national panel on educational behaviour has been only started today – we will have to wait twenty years from now for some results. This is exactly one of the strengths of social bookkeeping data: They have been collected for much longer and on a much wider range of topics than survey data. While internet log files exist only since the early 1990s and electronic company data bases since the mid-1960s, public administrational data have been collected in Germany since the 19th century.

Table 2: Sampling Problems Related to Specific Data Collection Mode

Data Collection Design	Timeslot for which data could be available in theory*	Missing Data / Information Gaps Could be Caused by		
		Data Selection I		Data Selection II
		Data Production (Cross-Sectional)	Data Production (Longitudinal)	Archiving
Retrospective Interviews	a human life-span (about 70 years back from time of data collection)	data production bias (selectivity of data producing institution, total nonresponse or missing values)		respondent might not remember past events (correctly)
Replication of Cross-Sectional Data	since the 1940s		danger of ecological fallacy, as it is impossible to analyze change on the level of single cases	depends on socio-cultural context of data production and archiving; multiplies over time due to: data getting lost or decaying due to neglect; purposeful destruction of data
Trend-Design	since the 1970s			
Panel-Design	since the 1980s		panel mortality spell-effects	
Process-Produced Mass Data	public administrational data since the 19th century; electronic company data bases since the mid-1960s internet log files since the early 1990s	depends on data type, e.g.: for public administrational data, forms could be empty; for newspapers, journalists could deem an event irrelevant	depends on data type, e.g.: for internet protocols, servers breakdown; for newspapers, changing importance of a topic for the general public	

* Note that there can be strong regional variation on the timeslots. For example, in the United states public administrational data were available much later, survey data earlier than in Germany.

Data Selection I: Data Production Bias

Choosing a data type is not only important because researchers have to find data at all. *Different data types are also affected in different ways by missing data, information gaps and data selection* (table 2). This becomes evident, of one compares survey data and social bookkeeping data (see *Hethey/Spengler* and *Huber/Schmucker* in this special issue) or social bookkeeping data from different sources (see *Köhler/Thomsen* and *Scioch/Oberschachtsiek* in this special issue).

For example, during data production, survey samples might be biased by the researcher's way of drafting the sample, by his inability to establish contact or by respondents refusal to participate (unit nonresponse) or reply to specific questions (missing values/item nonresponse). In contrast, the data lore discussed above illustrates that for social bookkeeping data, there are many reasons for case selection during data production which are strongly dependent on the data type. For example, public administrations might fail to establish first contact as the client may not want to disclose (certain) information. The administrational rules or a single clerk may not consider the client as a case worth collecting data on. Several organizations or clerks may interact in a way that data production is prevented or distorted (*Kruppe*, *Meyer* and *Köhler/Thomsen* in this special issue).

These examples also illustrate that *for mass data, measurement and sampling quality are closely entwined*.

When conducting longitudinal analysis, researchers may face the additional problem that *data producers may change their mode of data production over time*. That this in fact can strongly influence sample quality and statistical results, is illustrated by *Köhler/Thomsen*, *Seysen* and *Thorvaldsen* in this special issue.

The problem of handling sampling bias might become even more complex, if researchers want to *compare different populations* (Hage et al. 1980, Reuband 1980, Grunow 2004, Baur 2004). In this case, researchers have to decide

1) *which periods to compare with which other periods*. For example, of researchers want to compare employment careers, it makes a difference if they are interested in (a) a certain career phase, e.g. the mid-career. In this case, they can analyse people who were born at different periods in time, as long as data for the life period of interest are available. If one compares a person born 1940 with a person born 1960, the former entered the mid-career phase around 1970, the latter around 1990. (b) If researchers instead are interested how careers are affected by the same historical processes, e.g. unification, they need data for both persons for the same period of time, i.e. starting from 1990. (c) If researchers want to know how certain developments like modernization influence careers, they might need to analyse different periods for different countries. In this example the period would have to be analysed during which a country was modernized.
2) *if data are available for the relevant period.*
3) *if available data are comparable*. As *Tatjana Mika* points out in her paper, this might not be the case even for the same data type. For example, eligibility rules at public administrations might change the types of persons who can register at an administration or not.

In summary, (a) *the data production process itself will influences which cases are and which are not in the data base*, i.e. the equivalent to unit nonresponse in survey data. This point is discussed in this special issue by *Peter B. Meyer*, *Gunnar Thorvaldsen*, *Tatjana Mika*, *Markus Köhler* and *Ulrich Thomsen*, *Patrycja Scioch* and *Dirk Oberschachtsiek*, *Tanja Hethey* and *Anja Spengler*. (b) *The data production process may also result in missing data and inconsistencies,* i.e. the equivalent to item nonresponse in survey data. Markus Köhler and *Ulrich Thomsen*, *Patrycja Scioch* and *Dirk Oberschachtsiek*, *Tanja Hethey* and *Anja Spengler*, *Martina Huber* and A*lexandra Schmucker* think about some of the reasons for and results of these errors.

Data Selection II: Selection of Cases for Archiving

Even if data were originally produced, this does not mean that they will actually available for the researcher to analyse. Instead, human beings have to actually make an effort in order to preserve data. How precarious data preservation is, is demonstrated by Jagodzinski and Moschner (2008) and *Abrahamson et al.* (in this special issue) for standardized data and by Witzel et al. (2008) and *Christoph Thonfeld* (in this special issue) for qualitative data. Instead, it is much more likely that data will be lost due to neglect or natural decay or that data will be purposefully destroyed (see table 2). Moreover, which data are preserved and which are not strongly depends on the socio-cultural context which may change over time (Baur 2004, Baur/Lahusen 2005). At least since the 19th century, historians and public institutions have tried to preserve data by building archives where registrars systematically preserve data. However, there is only limited space in archives. Therefore, registrars usually archive only a sample of the data relevant for the archives topic. This means that in order for data to be archived, an archive has to exist, the archiving institution has to decide that the dataset is worth archiving. It still might decide not to archive specific cases or simply make mistakes during archiving. All these processes may further bias available data for future research and thus have to be assessed and handled by secondary researchers.

Data Selection III: Making Data Accessible for Researchers

While the sampling biases discussed so far all ranged around the topic of data being lost forever, another problem can be that data are never found again. From the point of view of a secondary researcher, the best data are useless, if she dies not know about the existence if a relevant archive or data set or is not granted access to the dataset or the set of variables relevant for her research. It thus has been a major point of discussion among archivists and data producers how to best make their data accessible to other researchers. Several papers in this special issue continue this debate: *Markus Köhler* and *Ulrich Thomsen* and *Christian Seysen* illustrate how complicated the processes are at the backend of a modern data producers and how many steps have to be taken in order to make mass data available to secondary researches. *Spyridoula Arathymou* explains recent efforts to develop for classical archives to make them more easily accessible for a general public. *George Alter*, *Kees Mandemakers* and *Myron Gutmann* suggest building an intermediate structure linking data suppliers with similar content and thus making them even more accessible for data users.

Data Selection IV: Selection of Cases for Analysis

Even of the secondary researcher gains access to all data relevant for his research field, he himself might select cases furthers. A first reason might be that processing cases might result in additional effort, e.g. if data still have to be coded. In these cases, drawing a random sample from the original data can make sense (Rohlinger 1982, Buchholz 2002). Additionally, researchers might purposefully delete cases from the data set for data analysis, e.g. they might concentrate on one specific sub-group or delete outliers. This issue is not discussed further in this special issue, as it sampling strategies for social bookkeeping data have been discussed recently in HSR by Buchholz (2002).

5. The Never-Ending Story: Archiving and Statistical Programmes

Mainstream social science has extensively discussed data collection and data analysis procedures. However, if one reads introductions to social research, it almost seems as nothing happens in between. In contrast, historians are typically strongly aware that in fact a lot of things do happen between those two phases of the research process and that a lot of things can go wrong during these intermediate research steps.

One of the reasons why sociologists tend to neglect these phases may be that a lot of issues are hard to systematize. Instead, due to their inherent nature, they probably always have to be discussed anew, as they (a) can only answered for each specific data type, (b) are affected by technological change or (c) by changing research questions.

How difficult it is to ensure archiving at all has been illustrated by several authors both for qualitative and quantitative data both in the past (Dollar 1980, Witzel et al. 2008; Jagodzinski/Moschner 2008) and in this issue of HSR (*Christoph Thonfeld*, *Mark Abrahamson et al.*)

Other issues arising during archiving are: How can the researcher or archiving institution transform historical sources into data bases? How should a database be designed and how should access to the database be organized? Which software packages are best suited for data management and data analysis?

Before the data can be archived and transformed into a data base, data have to undergo source criticism ("Quellenkritik", "Bewertungstheorie") (Müller

1982). Among the problems arising during this research phase are: coding open-ended and machine-readable data, reading standardized data into a data base, the differences between classical archives and digital archives and between critical editions and data bases (Buchmann 1980, Thaller 1986a-c, 1987, 1988, Härtel 1989, Panzeri 1989, Lipp 1989, Werner 1991, Jaritz 1991, Aumann et al 1999, Boonstra et al. 2004: 36-43, Volkens et al. 2009).[11]

Other problems that can arise when digitalizing historical data for analysis are orthography (i.e. spelling of names, persons and places). "An important decision is at which stage this standardization has to be carried out and how the coupling with the original information is preserved" (Boonstra et al. 2004: 44), i.e. how to de-contextualise not too much (Levermann 1991, Scheuermann 2006).

Furthermore, someone has to decide, if the data model should be structured around the logic of the sources, the logic of the registrars (*supply-side orientation*, see Boonstra 1990, Greenstein 1989) or the logic of researchers (*model-orientation*, Boonstra et al. 2004: 44-47).

As these issues are so complicated, *Data Managemen*t (Stone 1980, Pierau 2002) and *Historical Information Science* (Boonstra 1990, Boonstra et al. 2004) today are research fields in their own right.

A first debate in HSR on archiving techniques and types of software suitable for archiving and data analysis in quantitative historical research has been driven by Manfred Thaller. One subtheme of the debate were general principles of digital archives from the point of view of current technology (e.g. Reinke 1979, Stahlschmidt 1984, Thaller 1986a, 1987, 1988, 1991, Gathmann 1987, Sieglerschmidt 1988, Kapelle et al. 1988, Hänisch 1989, Greenstein 1989, Pasleau 1989, Trugenberger 1992, Huck 1993, Ofen 1993, Lenz 1993, Wettengel 1995, Melischek/Seethaler 1996, Aumann et al 1999, Pierau 2002).

Other papers introduced readers to specific data bases. They explained what type of data the data base contained and how the database was constructed (e.g. Irsigler 1978, Choppin 1989, Derosas 1989, Gippert 2002, Smets 1986, Sprengnagel 1987, Lloyd-Jones 1989, Meles 1989, Tidswell 1989, Ranieri 1989, Pierau 1991, Imfeld et al. 1995, Gabler/Steppe 1995, Rao/Marathe 1989, Van den Nieuwenhof, Patrick 2003). This discussion has moved to the HSR section "Data & Archives" in recent years.

Archiving problems are as pressing as ever. For example, *Thomas Kruppe* illustrates in his paper that data producing institutions may operationalize concepts differently from the researcher's needs. *Gunnar Thorvaldsen* points out

[11] Up to date and unknown to many social scientist, an own body of literature exists within history on this problem. Treffeisen (2004) has compiled a list of important references covering the following topics: source criticism, intermediate archives, archivists, parallel sources, mass sources, users' needs, special interests of various archives by archiving institution and topics.

that, furthermore, these concepts may change over time. As *Gunnar Thorvaldsen* and *Christian Seysen* show for Norway and Germany respectively, changes in mode of data collection and in technologies accompanying the data collection and archiving process may strongly influence data quality. How prone to errors and inconsistencies data are, is shown by *Tanja Hethey* and *Anja Spengler*, *Markus Köhler* and *Ulrich Thomsen, Martina Huber* and *Alexandra Schmucker*, *Patrycja Scioch* and *Dirk Oberschachtsiek*. All these authors illustrate how difficult it is for an archiving institution to build, organize and clean data for both storage and analysis. They also show that data are and have to be changed in order to be able to store them at all. If this process is not conducted carefully, it may be a source of error. Finally, it becomes also clear that all the procedures applied to data during the archiving process are strongly dependent on IT technology. As these technologies are prone to change, so will be the problems associated with them.

Even if an archive exists and is well organized, researchers face the problem of finding the right archive and finding the right data within the archive (Dollar 1980, Rowney 1986). Again, possibilities of access are at least technology-driven. *Spyridoula Arathymou* explains the current standards of documentation and archival description and aims at making archives more easily accessible to users. *George Alter*, *Kees Mandemakers* and *Myron Gutmann* go one step further: The suggest building an intermediate structure between suppliers and users in order to facilitate access to mass data and enable comparative research.

6. Data Preparation and Data Fusion (Record Linkage)

As can be seen from the discussion so far, errors and distortions are not the exception but normal when using social bookkeeping data. In comparison to survey research the question is therefore not how to avoid them but how to assess and handle them in research practice. Preparing and linking data therefore are much more important in mass data research than in survey research. As illustrated in graph 1 on page 17, data can be prepared during four steps of the research process: during data collection, when entering the archive, before being released from the archive and for actual data analysis. It is therefore no surprise, that data preparation has also received large attention in the methodological debate on mass data. As almost all articles on social science history, especially in HSR, cover data preparation issues at least to some extent, in this chapter I will give a short overview over important issues, link them to the issues of measurement, sampling and archiving discussed above and explain how the papers in this special issue relate to topics related to data preparation.

6.1 Data Preparation I: Preparing Data During Data Collection

Data often are prepared concurrently or immediately after data collection by the data producers. For example, a social security institution might check, if a client has entered wrong data and correct them, or it might process them for its own needs by entering information. Methodologically, these forms of data preparation can distort information and – as they are not controlled by registrars or researchers – are aspects of data production bias: They may cause both measurement and selection bias, but they may also decrease them. In order to assess these distortions, mass data have to undergo source criticism (Müller 1982) before being further processed. Although it is recommendable for the researcher to conduct source criticism herself, data suppliers increasingly conduct source criticism themselves before moving data to archives for social science research. In this special issue, *Gunnar Thorvaldsen*, *Peter B. Meyer* and *Tatjana* Mika give examples of what aspects source criticism could cover.

6.2 Data Preparation II & III: Preparing Data for Archiving & for Release From the Archive

Data can only be used by secondary researchers, if they have been archived after data collection and are later released again from the archive for research. While so far, the steps to be taken for survey and mass data research differed, from now on, there is no difference between handling survey and mass data research in principle:

When data enter an archive, they have to be prepared for the archive's special needs. A first problem to be solved can be transforming sources into data bases, especially if the original data are not machine-readable (Härtel 1989, Werner 1991, Jaritz 1991, Thaller 1986c).

Afterwards, registrars often further transform the data. Examples for such transformations are renaming variables; deleting cases or variables that are deemed unnecessary for archiving; deleting false information, transforming data into a common format, deleting information that for some reason cannot be archived (e.g. due to structure of the archive, due to archiving space). Archiving can also mean that information is added, e.g. documentation of the data generation process.

Before leaving an archive for secondary analysis, data are usually prepared again. Some of the reasons for data transformations are: deleting inconsistencies; deleting cases or variables the secondary researcher does not need or should for some reasons not process; preparing data in order to make it easier for the secondary researcher to use them.

That some *information is missing in the data or that data are inconsistent*, has been discussed both on survey and in mass data research methodology continuously (Rohlinger 1982, Best/Kuznia 1982, Schnell 1991, Rässler et al 2008). A major part of research conducted by data producing agencies directly addresses procedures for identifying and handling missing data and inconsistencies. In this special issue, *Christian Seysen*, *Tanja Hethey* and *Anja Spengler*, *Martina Huber* and *Alexandra Schmucker*, *Markus Köhler* and *Ulrich Thomsen*, *Gunnar Thorvaldsen*, *Thomas Kruppe*, *Patrycja Scioch* and *Dirk Oberschachtsiek* address different types of inconsistencies, explain ways of identifying and new procedures of deleting or at least handling them.

Another problem shared by survey and mass data research is that one single data set may not contain all the data needed for research and therefore different records have to be linked. However, they address the same problem from slightly different angles: In survey research, typically random samples are drawn. *Data Matching* (Rässler 2001, 2002), *Data Fusion* (IZ 2005, Rässler/Fleischer 1998, Rässler 2004, Kiesl/Rässler 2009) or *Data Integration* (IZ 2005) typically means linking two different small samples containing information on different persons. Therefore, information is not fused on the same person, but on two *different* persons who are statistically very similar. Still, this approach carries the danger of ecological fallacy. In contrast, mass data sets are usually much larger.

Therefore *Record Linkage* (Hershberg et al 1979, Hershberg 1980, Winchester 1980, Link 1999, Boonstra et al. 2004: 53-55) usually means combining data sets information on the *same* person. Several papers in this special issue discuss advanced problems of record linkage: *Christian Seysen*, *George Alter* et al., *Markus Köhler* and *Ulrich Thomsen* explain procedures for fusing data (a) across different measurement points in time and (b) across different data sets of the *same* data type. (c) Some newer developments aim at fusing different data sets of *different* data type, i.e. either different forms of social bookkeeping data (*Markus Köhler* and *Ulrich Thomsen*, *Patrycja Scioch* and *Dirk Oberschachtsiek*) or social bookkeeping data with survey data (*Tanja Hethey* and *Anja Spengler*, *Martina Huber* and *Alexandra Schmucker*).

Very often, data producers do not (only) provide microdata bot aggregated data. Data aggregation poses its own problems, as problems of microdata might be multiplied (Brosveet 1980, Vinovskis 1980). Therefore today, most data producers carefully document how data are aggregated, and these documents can be typically downloaded from their websites.

6.3 Data Preparation IV: Preparing Data During Data Analysis

Not only the archiving institution, also the secondary researcher often transforms data for her specific aims of data analysis. Typical examples are recoding variables, computing new variables and further reducing the sample to the group interesting for analysis. Some of the most common procedures are described in Baur and Fromm (2008b).

7. Open Questions

So what can be learned from the discussion so far? As we are still at the very beginning of methodological discussion and as there are still many open questions and problems unsolved, the ideal would be creating and consolidating a research field on mass data methodology, which could be similarly structured as survey methodology and mixed methods research. This research field could inquire which methodological problems are distinct for mass data, which problems mass data do share with other process-produced data and/or survey methodology and how different data types can be mixed. In section 7.2, I suggest some possible fields of research. While a mass data research would structure communication between social science methodologists, a second problem is transferring knowledge – from one generation of researchers to the next, and from methodologists to students and researchers using public administrational data.

7.1 Knowledge Transfer

The best methodological procedure is of no use, if no one applies it. Therefore knowledge transfer has been and will have to be a primary goal of mass data methodology. As stated in the introduction, there is a danger of knowledge loss due to generational change and shifting research fields. We hope that with this special issue, we have moved one step towards saving existing knowledge.

Even if knowledge transfer is secured on the level of methodologically competent researchers, the problem of transferring methodological knowledge to people needing this knowledge but not yet having it or even being unaware of needing it still remains unsolved. For mass data, this concerns at least two groups of persons:

1) *Supply-Side Training:* All research on mass data and secondary analysis of survey data shows that measurement and selection bias can only be assessed and properly handled, if researchers are knowledgeable about the data production process. The better data are documented, the easier handling them properly is. Thus, one of the first discussions in the 1980s was the question of how to teach *data producers* how to document their data (Nielson 1980, Kropac 1986; Hall 1989). As Abrahamson et al. (in this special issue) show, this is still a problem for survey data even in countries like the U.S. In Germany, GESIS[12] has established a structure which helps primary researchers to document and archive their data (although it is not as widely used as it could be), and researchers from ZHSF[13] have trained archiviarians and data

[12] "GESIS – Leibniz Institute for the Social Sciences", www.gesis.org.

[13] "Zentrum für Historische Sozialforschung", today integrated into GESIS.

producing institutions in the last three decades in order to ensure standards of documentation.

2) *User-Training:* Additionally, *students and researchers* new to the field need to be trained in doing quantitative historical analysis.

The 1980s discussion on user-training started with general discussions on what students and researchers needed to know in order to do quantitative historical analyses and what a basis curriculum on historical social research should look like (Best/Schröder 1981, Bauer 1982, Ayton 1989). Soon, it was clear that there war a defined set of problems: Students and researchers using mass data need to know how to use a computer (Rowney 1986, Trainor 1987, Hirschheim et al. 1988). They need to find and be able to use the right archives and the data bases within the archive (Schurer 1986, 2006, Pierau 2002). They need to be able to assess the quality of the sources, to apply statistics and to use the statistical packages.

As a result of this discussion, within the context of HSR, there was a discussion (driven by Manfred Thaller) on the advantages of various software packages used in quantitative historical analysis, and some of the HSR supplements are introductory books to data management and statistics (Best/Mann 1977, Thome 1989/1990, Andreß 1990, Best/Thome 1992, Sensch 1995, Rahlf 1998, Pierau 2002). As source quality can only be properly assessed and handled in relation to a specific research question, Schröder et al. (2000) developed a course programme teaching the steps to be taken during the whole research process, using the example of parliaments, elites and biographies.

In order to facilitate access to public administrational data, at least in Germany, most data producing institutions have established Research Data Centres[14] during the last ten years. A lot of documents on data production can be obtained from the data producing institutions themselves. E.g., complete documentation on how federal statistics are produced, can be found at the Office for Federal Statistics.[15] Recently, the RatSWD has been created as an umbrella organization.[16]

[14] "Research Data Centers make individual data accessible for scientific research by and large through the creation of factually anonymized data sets (Scientific Use Files) released to research institutions. In exceptional cases, where research concerns particularly sensitive data, or where it is not possible to adequately anonymize data without the loss of information, data access is possible through the creation of workplaces for guest researchers at specific Research Data Centers, or through the development of a system for controlled remote data access" (www.ratswd.de).

[15] An overview over data-producing institutions can be found in Baur/Fromm (2008a).

[16] "The Council's main purpose is to advise in the development of the German data infrastructure for empirical research in the social and economic sciences. The Council is working not only to increase access to microdata and to sustainably improve data quality, but increasingly also in the development of long-term data surveys, together with both official government institutions (official statistical offices, social insurance institutions, government

While these structures facilitate data access also for experienced users, access for new users is facilitated by special workshops and training courses showing how to use data. Additionally, GESIS offers – especially with the ZHSF Autumn Seminars ("ZHSF-Herbstseminar", "ZHSF-Methodenseminar") – special courses in training both statistical procedures and in applying mass data methodology.

This model is very effective in knowledge transfer, and today, many German researchers are using process-produced mass data. It could well serve as a model of organising knowledge transfer in other countries. Ironically, this model is in danger of becoming a victim of its own success: Exactly because it is working so well, there is currently the danger of cutting budgets, as younger researchers and politicians are not aware any more of how important these structures are for knowledge transfer.

7.2 Reopening the Methodological Discussion

While knowledge transfer to suppliers and new users has been relatively successful, a field of research on mass data methodology still has to be established. The aims of such a research field have already been stated by Bick and Müller in the 1980s: Equivalent to research on survey methodology, there should be research on distortions of process-produced data during the data production and archiving process.

Very likely, s*ome issues will always have to be discussed anew*, as they (a) can only answered for each specific data type, (b) are affected by technological change or (c) by changing research questions. Among these problems are: How should a database be designed, and how should access to the database be organized? Which software packages are best suited for data management and data analysis?

However, for other questions, the possibility of *accumulation of methodological knowledge* at least exists. From a *source-perspective*, such an amassing of knowledge might be possible concerning the individual characteristics of specific data types and sources, leading to a *data lore ("Datenkunde")*, as many data producers continue collecting data in the same way for long times and as clients adopt typical behaviours to data collecting institutions. While much has already been said about e.g. criminological and census data (see compilation of older referenced in HSR Trans 22), there are fields of research and data sources which are relatively new. For the latter, a data lore still has to be developed. Examples for new data types are internet data and companies'

research units, etc.) and non-governmental institutions (universities and non-university research institutes, e.g., Leibniz Society institutions). All of the Council's work is fundamentally based on the objective of fostering constructive dialogue between the research community and data production facilities" (www.ratswd.de).

customer databases. Examples for newer research fields are research on the media, organizations and markets.

This source-perspective should be complemented by a *research-process-perspective* which discusses typical problems arising from mass data along the research problems along the research process, enhancing process-quality. As can be seen from the structure of the papers in this special issue, authors still struggle for a format for discussion mass data methodology. The papers also illustrate how difficult it is not to think in terms of content but in terms of more general methodological problems. Still, the papers indicate where the discussion could lead:

How and why are mass data biased (both concerning measurement and sample quality)? Who causes these distortions, and at which stages of the research process do they arise? How large are these effects? Are distortions specific to certain data types, countries, institutions and/or historical periods, or are there general patterns, i.e. can we predict distortions and bias? Can they at least partly be avoided, e.g. by better training of data suppliers? How could and should these distortions be handled during data preparation?

It would also make sense to lead the discussion cross-culturally from the beginning and also systematically link it to survey methodology and mixed methods research from the beginning. In the long run, the ideal would be to integrate results in introductions to social research.

References

Abbott, Andrew (2001): Time Matters. On Theory and Method. Chicago/London: Chicago University Press.

ADM (Ed.) (2004): Nonresponse und Stichprobenqualität. Frankfurt a.M.: Verlagsgruppe Deutscher Fachverlag.

Andreß, Hans-Jürgen (1992): Einführung in die Verlaufsdatenanalyse. Historical Social Research Supplement 4.

Aumann, Stefan/Ebeling, Hans-Heinrich/Fricke, Hans-Reinhard/Hoheisel, Peter/Rehbein, Malte/Thaller, Manfred (1999): From Digital Archive to Digital Edition. In: Historical Social Research 24 (1). 101-144.

Ayton, Andrew (1989): Computing for History Undergraduates: A Strategy for Database Integration. In: Historical Social Research 14 (4). 46-51.

Bartholomew, David J (Ed.) (2006): Measurement. 4 Volumes. London/New Delhi/Newbury Park: Sage.

Bauer, Henning (1982): Der Einsatz archivierter Daten in der Lehre der Historischen Sozialforschung. In: Historical Social Research 7 (4). 63-72.

Baur, Nina (2004): Wo liegen die Grenzen quantitativer Längsschnittsanalysen? Series: Bamberger Beiträge zur empirischen Sozialforschung 23. Bamberg.

Baur, Nina (2005): Verlaufsmusteranalyse. Methodologische Konsequenzen der Zeitlichkeit sozialen Handelns. Wiesbaden: VS-Verlag.

Baur, Nina (Ed.) (2009a): Linking Theory and Data: Process-Generated and Longitudinal Data for Analysing Long-Term Social Processes. Special Issue of Historical Social Research 34 (1).

Baur, Nina (2009b): Problems of Linking Theory and Data in Historical Sociology and Longitudinal Research. In: Historical Social Research 34 (1). 7-21.

Baur, Nina (2009c): Memory and Data. Methodological Implications of Ignoring Narratives in Survey Research. In: Packard, Noel (Ed.) (2009): Sociology of Memory. In Print.

Baur, Nina/Fromm, Sabine (2008a): Nützliche Software und Fundorte für Daten. In: Baur, Nina/Fromm, Sabine (Eds.) (2008b): Datenanalyse mit SPSS für Fortgeschrittene. Wiesbaden: VS-Verlag. 208-214.

Baur, Nina/Fromm, Sabine (Eds.) (2008b): Datenanalyse mit SPSS für Fortgeschrittene. Wiesbaden: VS-Verlag.

Baur, Nina/Lahusen, Christian (2005): Sampling Process-Generated Data in Order to Trace Social Change: The Case of Newspapers. In: van Dijkum, Cor/Blasius, Jörg/Durand, C. (Eds.) (2005): Recent Developments and Applications in Social Research Methodology. Opladen: Barbara Budrich.

Baur, Nina/Lamnek, Siegfried (2007): Multivariate Analysis. In: Ritzer, George (Ed.): The Blackwell Encyclopedia of Sociology. Blackwell Publishing Ltd. 5176-5179.

Behnke, Joachim/Behnke, Nathalie/Baur, Nina (2006): Empirische Methoden der Politikwissenschaft. Paderborn: Ferdinand Schöningh.

Best, Heinrich (1988): Historische Sozialforschung als Erweiterung der Soziologie. In: KZfSS 40 (1). 1-15. Reprinted in: Historical Social Research Supplement 20 (2008). 74-89.

Best, Heinrich (1996): Historische Sozialforschung und Soziologie. Reminiszenzen und Reflektionen zum zwanzigsten Jahrestag der Gründung der Arbeitsgemeinschaft QUANTUM. In: Historical Social Research 21 (2). 81-90. Reprinted in: Historical Social Research Supplement 20 (2008). 90-99.

Best, Heinrich (2008a): Historische Sozialforschung und Soziologie. Reminiszenzen und Reflektionen zum zwanzigsten Jahrestag der Gründung der Arbeitsgemeinschaft QUANTUM. In: Historical Social Research Supplement 20. 90-102.

Best, Heinrich (2008b): Quantifizierende Historische Sozialforschung in der Bundesrepublik Deutschland. Ein Überblick. In: Historical Social Research Supplement 20. 49-73.

Best, Heinrich/Kuznia, Reiner (1983): Die Behandlung fehlender Werte bei der seriellen Analyse namentlicher Abstimmungen oder: Wege zur Therapie des Horror Vacui. In: Historical Social Research 8 (2). 49-82.

Best, Heinrich/Mann, Reinhard (Eds.) (1977): Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 3. Stuttgart: Klett-Cotta.

Best, Heinrich/Schröder, Wilhelm Heinz (1981): Basiscurriculum für eine Quantitative Historische Sozialforschung: Vorschläge für eine Einführungsveranstaltung am Beispiel des Zentrum-Herbstseminars. In: Historical Social Research 6 (1). 3-50.

Best, Heinrich/Schröder, Wilhelm Heinz (1987): Quantitative Historical Social Research. The German Experience. In: Jarausch, Konrad H./Schröder, Wilhelm Heinz (Eds.): Qualitative History of Society and Economy. Some International

Studies. Historisch-Sozialwissenschaftliche Forschungen Volume 21 St. Katharinen: Scripta Mercaturae. 30-48. Reprinted in: Schröder, Wilhelm Heinz (Ed.) (2006): Historisch-Sozialwissenschaftliche Forschungen: Quantitative sozialwissenschaftliche Analysen von historischen und prozeß-produzierten Daten. Historical Social Research Supplement 18. 120-135.

Best, Heinrich/Thome, H. (Eds.) (1992): Neue Methoden der Analyse historischer Daten. Historisch-Sozialwissenschaftliche Forschungen Volume 23. St. Katharinen: Scripta Mercaturae.

Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Eds.) (1984a): Sozialforschung und Verwaltungsdaten. Historisch-Sozialwissenschaftliche Forschungen Volume 17. Stuttgart: Klett-Cotta.

Bick, Wolfgang/Müller, Paul J. (1977): Buchführung der Verwaltungen als sozialwissenschaftliche Datenbasis. In: Müller, Paul J. (Ed.) : Die Analyse prozeß-produzierter Daten. Historisch-Sozialwissenschaftliche Forschungen Volume 2. Stuttgart: Klett-Cotta. 42-88.

Bick, Wolfgang/Müller, Paul J. (1980): The Nature of Process-Produced Data. Towards a Social-Scientific Source Criticism. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 369-413.

Bick, Wolfgang/Müller, Paul J. (1984): Sozialwissenschaftliche Datenkunde für prozeßproduzierte Daten. Entstehungsbedingungen und Indikatorenqualität. In: Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Eds.) (1984a): Sozialforschung und Verwaltungsdaten. Historisch-Sozialwissenschaftliche Forschungen Volume 17. Stuttgart: Klett-Cotta. 123-159. Reprinted as a short version in: Historical Social Research 27 (2/3). 227-252.

Biemer, Paul B./Groves, Robert M./Lyberg, Lars E. (2004): Measurement Errors In Surveys. New York: John Wiley & Sons Inc.

Biemer, Paul/Lyberg, Lars E. (2003): Introduction to Survey Quality. Hoboken: John Wiley & Sons.

Biste, Bärbel/Hohls, Rüdiger (Eds.) (2000): Fachinformation und EDV. Arbeitstechniken für Historiker. Einführung und Arbeitsbuch. Historical Social Research Supplement 12.

Boonstra, Onno W. (1990): A. Supply-Side Historical Information Systems. The Use of Historical Databases in a Public Record Office. In: Historical Social Research 15 (1). 66-71.

Boonstra, Onno/Breure, Leen/Doorn, Peter (2004): Past, Present and Future of Historical Information Science. In: Historical Social Research 29 (2) 4-132.

Brehm, John (1993): The Phantom Respondents. Ann Arbour: The University of Michigan Press.

Brosveet, Jarle (1980): Standardization of Longitudinal, Aggregate-level Data in the Norwegian Commune Database. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 513-523.

Buchholz, Matthias (2002): Stichprobenverfahren bei massenhaft gleichförmigen Einzelfallakten. Eine Fallstudie am Beispiel von Sozialhilfeakten. In: Historical Social Research 27 (2/3). 100-223.

Buchmann, Wolf (1980): Archives and Machine-readable Data from Public Administration in the Federal Republic of Germany. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 473-476.

Bulmer Martin (Ed.) (2004): Questionnaires. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Bulmer Martin (Ed.) (2010): Questionnaires 2. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Bulmer Martin/Sturgis Patrick J./Allum, Nick (Ed.) (2009): The Secondary Analysis of Survey Data. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Bulmer, Martin (Ed.) (2003): Questionnaires. 4 Volumes. London/New Delhi/Newbury Park, Sage Publications.

Burkhardt, Martin (2006): Arbeiten im Archiv. Praktischer Leitfaden für Historiker und andere Nutzer. Paderborn et al: Schöningh.

Campbell, Donald T./Russo, M. Jean (2001): Social Measurement. London/New Delhi/Newbury Park: Sage Publications.

Choppin, Alain (1989): EMMANUELLE: A Data Base for Textbooks' History in Europe. In: Historical Social Research 14 (4). 52-58.

Clubb, Jerome M./Scheuch, Erwin K. (Eds.) (1980): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta.

Creswell, John W./Plano Clark, Vicki (2006): Designing and Conducting Mixed Methods Research. London et al.: Sage.

Creswell, John W./Plano Clark, Vicki/Garrett, Amanda L. (2008): Methodological Issues in Conducting Mixed Methods Research Designs. In: Bergmann, Manfred Max (2008): Advances in Mixed Methods Research. Lost Angeles et al.: Sage. 66-83.

De Vaus, Davis (Ed.) (2002): Social Surveys. 2 Volumes. London/Thousand Oaks/New Delhi: Sage.

De Vaus, Davis (Ed.) (2007): Social Surveys 2. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Derosas, Renzo (1989): A Database for the Study of the Italian Population Registers. In: Historical Social Research 14 (4). 59-65.

Diekmann, Andreas (2006): Empirische Sozialforschung. Reinbek: Rowohlt.

Dillman, Don A. (2007): Mail And Internet Surveys: The Tailored Design Method. 2nd Edition. New York: John Wiley & Sons Inc.

Dollar, Charles M. (1980): Problems and Procedures for Preservation and Dissemination of Computer-readable Data. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 457-472.

Durkheim, Émile (1897): Le Suicide. Paris: Félix Alcan.

Esposito, James L./Rothgeb, Jennifer M. (1997): Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In: Lyberg, Lars et. al. (Ed.) (1997): Survey Measurement and Process Quality. New York u. a.: John Wiley & Sons. 541-572.

Fink, Arlene (2003): The Survey Kit. London/New Delhi/Newbury Park: Sage.

Fuchs, Marek (1994): Umfrageforschung mit Telefon und Computer. Weinheim: Beltz.

Gabler, Hans Walter/Steppe, Wolfhard (1995): Shakespeare-Datenbank. Auf der Grundlage der Englisch-deutschen Studienausgabe der Dramen Shakespeares. In: Historical Social Research 20 (4). 155-159.

Gathmann, Immo (1987): Datenbank und Informationsverwaltungssysteme. Probleme ihrer Implementierung auf nur einmal beschreibbaren Speichermedien. In: Historical Social Research 12 (1). 76-87.

Gelanty, Gerard/Isin, Egin F. (Eds.) (2003): Handbook of Historical Sociology. London/Thousand Oaks/New Delhi: Sage.

Gippert, Jost (2002): Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems. In: Historical Social Research 27 (1). 207-214.

Greenstein, Daniel I. (1989): A Source-Oriented Approach to History and Computing: The Relational Database. In: Historical Social Research 14 (3). 9-16.

Groves, Robert M. (2004): Survery Errors and Survey Costs. Hoboken: John Wiley & Sons.

Groves, Robert M./Eltinge, John L./Little, Roderick (2001): Survey Nonresponse New York: John Wiley & Sons Inc.

Groves, Robert M./Fowler, Floyd J./Couper, Mick P. (2004): Survey Methodology. New York: Wiley-Interscience.

Grunow, Daniela (2004): The Dynamics of Gender and Social Change: A Comparison of Male and Female Employment Careers in West Germany, East Germany and Denmark. Präsentation im Rahmen des Doktorandenkolloquiums „Lebensverläufe im Globalisierungsprozess". Otto-Friedrich-Universität Bamberg. 2.3.2004.

Hage, Jerald/Gargan, Edward T./Hannemann, Robert (1980): Procedures for Periodizing History. Determining Eras in the Histories of Britain, France, Germany and Italy. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 267-283.

Hall, John A./Bryant, Joseph M. (Eds.) (2005): Historical Methods in the Social Sciences. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Hall, John R. (2007): Historicity and Sociohistorical Research. In: Outhwaite, William/Turner, Stephen P. (Eds.): The SAGE Handbook of Social Science Methodology. London: SAGE. 82-99.

Hall, Ninette van (1989): Towards A Standard for the Description of Historical Datasets. In: Historical Social Research 14 (1). 89-117.

Hänisch, Dirk (1989): Inhalt und Struktur der Datenbank „Wahl- und Sozialdaten der Kreise und Gemeinden des Deutschen Reiches von 1920 bis 1933". In: Historical Social Research 14 (1). 39-67.

Härtel, Reinhard (1989): To Treat or not to Treat: the Historical Source Before the Input. In: Historical Social Research 14 (1). 25-38.

Hershberg, Theodore (1980): Interdisciplinary Research at the Philadelphia Social History Project. Analytic Goals, Data and Data Manipulation Strategies for the Study of the Nineteenth-Century Industrial City. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 84-111.

Hershberg, Theodore/Burstein, Alan/Dockhorn, Robert (1979): Verkettung von Daten. Record Linkage am Beispiel des Philadelphia Social History Project. In:

Schröder, Wilhelm Heinz (Ed.): Moderne Stadtgeschichte. Historisch-Sozialwissenschaftliche Forschungen Volume 8. Stuttgart: Klett Cotta. 35-73.

Hirschheim, R./Smithon, S./Whitehouse, D. (1988): Microcomputer Use in the Humanities and Social Sciences: A United Kingdom Survey. In: Historical Social Research 13 (2). 141-144.

Hoffmeyer-Zlotnik, Jürgen H.P./Harkness, Janet (Eds.) (2005): Methodological Aspects in Cross-National Research, Mannheim. Mannheim: ZUMA.

Hoffmeyer-Zlotnik, Jürgen H.P./Wolf, Christof (Eds.) (2003): Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer Academic Publishers.

Hox, Joop J. (1997): From Theoretical Concept to Survey Question. In: Lyberg, Lars et. al. (Eds.) (1997): Survey Measurement and Process Quality. New York u. a.: John Wiley & Sons. 45-69.

Huck, Thomas-Sergej (1993): Einsatzmöglichkeiten elektronischer Datenbanken in der Geschichtswissenschaft am Beispiel einer Untersuchung über das Zisterzienserkloster Hardehausen (1140-1803). In: Historical Social Research 18 (1). 71-91.

Hunter, A./ Brewer, J. (2003): Multimethod Research in Sociology. In: Tashakkori, A./Teddlie, C. (Eds.): Handbook of Mixed Methods in Social and Behavioral Research. Thousand Oaks et al.: Sage. 577-594.

Imfeld, Klaus/Pfister, Christian/Häberli, Peter (1995): BERNHIST: Eine raumzeitliche Datenbank für den Schweizer Kanton Bern im Internet. In: Historical Social Research 20 (3). 102-133.

Irsigler, Franz (Ed.) (1978): Quantitative Methoden in der Wirtschafts- und Sozialgeschichte der Vorneuzeit. Historisch-Sozialwissenschaftliche Forschungen Volume 4. Stuttgart: Klett-Cotta.

IZ (Informationszentrum Sozialwissenschaften) (Ed.) (2005): Datenfusion und Datenintegration. 6. Wissenschaftliche Tagung des IZ. Sozialwissenschaftliche Tagungsberichte 10. Bonn: IZ.

Jagodzinski, Wolfgang/Moschner, Meinhard (2008): Archiving Poll Data. In: Donsbach, Wolfgang/Traugott, Michael W. (Eds.) (2008): Public Opinion Research. Los Angeles et al.: Sage. 468-476.

Jarausch, Konrad H. (1987): (Inter)national Styles of Quantitative History. In: Jarausch, Konrad H./Schröder, Wilhelm Heinz (Eds.): Qualitative History of Society and Economy. Some International Studies. Historisch-Sozialwissenschaftliche Forschungen Volume 21. St. Katharinen: Scripta Mercaturae. 5-18.

Jaritz, Gerhard (1991): The Image As Historical Source Or: Grabbing Contexts. In: Historical Social Research 16 (4). 100-105.

Johnson, R. B./Turner, L. A. (2003): Data Collection Strategies in Mixed Methods Research. In: Tashakkori, A./Teddlie, C. (Eds.): Handbook of Mixed Methods in Social and Behavioral Research. Thousand Oaks et al.: Sage, 297-320.

Kapelle, Günther/Reymann, Wolfgang/Schwarz, Rainer (1988): Aufbau und Analyse der Datenbank „Sozialgeschichte Berlins von 1650 bis 1799“. In: Historical Social Research 13 (4). 3-54.

Kiesl, Hans/Rässler, Susanne (2009): How Valid Can Data Fusion Be? In Print in: Journal of Official Statistics.

King, Gary/Keohane, Robert O./Verba, Sidney (1994): Designing Social Inquiry. Princeton: Princeton University Press.

Koch, Achim/Porst, Rolf (Eds.) (1998): Nonresponse in Survey Research. ZUMA-Nachrichten Spezial 4: Mannheim, ZUMA.

Kromrey, Helmut (2006): Empirische Sozialforschung. Opladen: Leske + Budrich.

Kropac, Ingo (1986): Von der Quelle zum Datensatz. In: Thaller, Manfred (Ed.): Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 20. St. Katharinen: Scripta Mercaturae. 199-216.

Lenz, Hans-J. (1993): On the Design of a Statistical Database, Micro-, Macro- and Metadata Modelling. In: Historical Social Research 18 (4). 31-48.

Levermann, Wolfgang (1991): Historical Data Bases and the Context Sensitive Handling of Data. Towards the Development of Historical Data Base Management Software. In: Historical Social Research 16 (4). 74-88.

Lipp, Carola (1990): Symbolic Dimensions of Serial Sources. Hermeneutical Problems of Reconstructing Political Biographies Based on Computerized Record Linkage. In: Historical Social Research 15 (1). 30-40.

Lippe, Peter von der (1998): Die amtliche Statistik der DDR: „Fälschungen“ oder „spezifische Form der Manipulation, zentral vollzogen“? In: Historical Social Research 23 (1/2). 339-343.

Lloyd-Jones, Roger (1989): Manchester: A Database. In: HSR 14 (3). 35-41.

Ludwig-Mayerhofer, Wolfgang (2003): Zur Qualität der sozialwissenschaftlichen Methodenausbildung – am Beispiel statistischer Datenanalyse. In: ZA-Informationen 53. 144-155.

Mayer, Thomas (2009): Wie kommt die Eugenik in die Eugenik? Sampling und Auswahlverfahren von prozessproduzierten Daten am Beispiel eugenischer Netzwerke in Österreich. In: Historical Social Research 34 (1). 159-171.

Meles, Brigitte (1989): The Swiss Database Project for Art and Cultural Heritage. In: Historical Social Research 14 (3). 94-97.

Melischek, Gabriele/Seethaler, Josef (1996): Konzept und Anforderungen einer Institutional Process Analysis am Beispiel der Datenbank der Wiener Tageszeitungen (1918-1938) Datenbank der Wiener Tageszeitungen (1918-1938). In: Historical Social Research 21 (3). 57-75.

Müller, Paul J. (Ed.) (1977): Die Analyse prozeß-produzierter Daten. Historisch-Sozialwissenschaftliche Forschungen Volume 2. Stuttgart: Klett-Cotta.

Müller, Paul J. (1982): Improving Source Criticism to Cope with New Types of Sources and Old Ones Better. In: Historical Social Research 7 (4). 25-33.

Nielson, Per (1980): How to Teach Data Producers “The Noble Art” of Data Documentation. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 477-487.

Ofen, Ulrich von (1993): Relational Database-Structures in Archaeology: ADS – an Application in Client-Server-Conception Developed by Means of CASE*Methode. In: Historical Social Research 18 (3). 22-34.

Panzeri, Matteo (1989): Automatic Indexes of Literary Sources for Art History. The Notizie by Frederico Alizeri. In: Historical Social Research 14 (4). 10-16.

Pasleau, Suzy (1989): Historical Data Bases as a Field for Structured Query Language. In: Historical Social Research 14 (1989): (3). 23-29.

Pierau, Karl (1991): Ein konzeptuelles Schema für Familiendaten. In: Historical Social Research 16 (1). 48-59.

Pierau, Karl (2002): Datenbank- und Informationsmanagement in der Historischen Sozialforschung. Eine praxisorientierte Einführung. Historical Social Research Supplement 14.

Platt, Jennifer (1981): Evidence and Proof in Documentary Research. In: Sociological Review 29 (1). 31-66. Reprinted in: Hall, John A./Bryant, Joseph M. (Eds.) (2005): Historical Methods ins the Social Sciences. Volume III: The Logic of Historical Sociological Inquiry. London/Thousand Oaks/New Delhi: Sage. 165-198.

Presser, Stanley/Rothgeb, Jennifer M./Couper, Michael P. (2004): Methods for Testing and Evaluating Survey Questionnaires. New York: John Wiley & Sons.

Prüfer, Peter (1996): Verfahren zur Evaluation von Survey-Fragen. Ein Überblick. In: ZUMA-Nachrichten 20 (39). 95-116.

Rahlf, Thomas (1998): Deskription und Inferenz. HSR-Supplement 9.

Ranieri, Fillipo (1986): Eine Datenbank über juristische Dissertationen und Juristen im Alten Reich. Ein Projektbericht. In: Historical Social Research 11 (1). 109-115.

Ranieri, Fillipo (1989): The Lawyers in the Holy Roman Empire of the 16th to the 18th Century. A Historical Data Base. In: Historical Social Research 14 (3). 62-67.

Rao, Mythili/Marathe, Ashok (1989): South Indian Megalithic Culture: Database and its Application. In: Historical Social Research 14 (4). 17-23.

Rässler, Susanne (2001): Alternative Approaches to Statistical Matching with an Application to Media Data. Habilitationsschrift Nürnberg.

Rässler, Susanne (2002): Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Lecture Notes in Statistics 168. New York: Springer.

Rässler, Susanne (2004): Data Fusion: Identification Problems, Validity, and Multiple Imputation. In: Austrian Journal of Statistics 33. 153-171.

Rässler, Susanne, Fleischer, K. (1998): Aspects Concerning Data Fusion Techniques. In: ZUMA Nachrichten Spezial. Mannheim. 317-333.

Rässler, Susanne/Rubin, D.B./Schenker, N. (2008): Incomplete Data: Diagnosis, Imputation, and Estimation. In: de Leeuw, E.D./Hox, J.J./Dillman, D.A. (Eds.): International Handbook of Survey Methodology. Hillsdale: Lawrence Erlbaum Associates. 370-386.

Reinke, Herbert (1979): Archiving Machine-Readable Historical Data: Data Services of the Center for Historical Social Research. In: Historical Social Research 4 (4). 36-38.

Reuband, Karl-Heinz (1980): Life Histories. Problems and Prospects of Longitudinal Designs. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 135-163.

Roberts, Caroline (Ed.) (2008): Attitude Measurement. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Rohlinger, Harald (1982): Quellen als Auswahl – Auswahl aus Quellen. In: Historical Social Research 7 (4). 34-62.

Rokkan, Stein (1976): Data Services in Western Europe. Reflections and Variations in the Conditions of Academic Institution Building. In: Amercian Behavioural Scientist 19. 443-454.

Rost, Jürgen (2004): Lehrbuch Testtheorie/Testkonstruktion. Bern, Huber.

Rowney, Don Karl (1986): The Microcomputer in Historical Research. Accessing Commercial Databases. In: Thaller, Manfred (Ed.): Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 20. St. Katharinen: Scripta Mercaturae. 217-231.

Ruloff, Dieter (1985): Historische Sozialforschung. Einführung und Überblick. Stuttgart.

Salheiser, Axel (2009): Handling Ideological Bias and Shifting Validity of Longitudinal Data: The Case of Process-Generated Data on GDR Elites. In: Historical Social Research 34 (1). 197-210.

Salkind, Neil J. (Ed.) (2006): Encyclopedia of Measurement and Statistics. 4 Volumes. London/New Delhi/Newbury Park: Sage Publications.

Scheuch, Erwin (1977): Die wechselnde Datenbasis der Soziologie. Zur Interaktion zwischen Theorie und Empirie. In: Müller, Paul J. (Ed.): Die Analyse prozeß-produzierter Daten. Historisch-Sozialwissenschaftliche Forschungen Volume 2. Stuttgart: Klett-Cotta. 5-41. Reprinted in: Schröder, Wilhelm Heinz (Ed.) (2006): Historisch-Sozialwissenschaftliche Forschungen: Quantitative sozialwissenschaftliche Analysen von historischen und prozeß-produzierten Daten. Historical Social Research Supplement 18. 24-46.

Scheuermann, Leif (2006): Ontologien in den historischen Wissenschaften In: Historical Social Research 31 (3). 308-316.

Schnell, Rainer (1991): Realisierung von Missing-Data-Ersatztechniken innerhalb statistischer Programmpakete und ihre Leistungsfähigkeit. In: HSR 23. 105-137.

Schnell, Rainer (2000): Nonresponse in Bevölkerungsumfragen. Opladen: Leske+ Budrich.

Schnell, Rainer/Hill, Paul B./Esser, Elke (2005): Methoden der empirischen Sozialforschung. München/Wien: R. Oldenbourg.

Schröder, Wilhelm Heinz (1988): Historische Sozialforschung. Forschungsstrategie, Infrastruktur, Auswahlbibliographie. Historical Social Research Supplement 1.

Schröder, Wilhelm Heinz (1994): Historische Sozialforschung. Identifikation, Organisation, Institution. Historical Social Research Supplement 6.

Schröder Wilhelm Heinz/Weege, Wilhelm/Zech, Martina (2000): Historische Parlamentarismus-, Eliten- und Biographieforschung. Historical Social Research Supplement 11.

Schurer, Kevin (1986): Historic Data Bases and the Researcher. In: Thaller, Manfred (Ed.): Datenbanken und Datenverwaltungssysteme als Werzeuge historischer Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 20. St. Katharinen: Scripta Mercaturae. 145-157.

Schützeichel, Rainer (2004): Historische Soziologie. Transcript.

Scott, Jacqueline/Xie, Yu (Eds.) (2005): Quantitative Social Sciences. 4 Volumes. London/Thousand Oaks/New Delhi: Sage.

Scott, John P. (Ed.) (2006): Documentary Research. 4 Volumes. London/ Thousand Oaks/New Delhi: Sage.

Sensch, Jürgen (1995): Statistische Modelle in der historischen Sozialforschung. Historical Social Research Supplement 7.

Sieglerschmidt, Jörn (1988): Probleme des Aufbaus und Umfangs einer Datenbank „Historische Statistik“. In: Historical Social Research 13 (1). 89-110.

Smets, Josef (1986): The South-French Society and the French Revolution – The Creation of a Great Data Base with CLIO. In: Historical Social Research 11 (2). 96-105.

Spohn, Willfried (2000): Historische Soziologie. In: Münch, Richard / Jauß, Claudia / Stark, Carsten (Eds.) (2000): Soziologie 2000. Kritische Bestandsaufnahmen zu einer Soziologie für das 21. Jahrhundert. Soziologische Revue. Sonderheft 5. 101-116.

Sprengnagel, Gerhard (1987): „Wiener Neustadt im Industriezeitalter“: Eine Datenbank zur Sozialgeschichte einer österreichischen Industrieregion im 19. Jahrhundert. In: Historical Social Research 12 (1). 3-27.

Stahlschmidt, Rainer (1984): Ein Archivierungsprogramm für das Datenmaterial der amtlichen Statistik. In: Bick, Wolfgang/ Mann, Reinhard/ Müller, Paul J. (Eds.): Sozialforschung und Verwaltungsdaten. Historisch-Sozialwissenschaftliche Forschungen 17. Stuttgart: Klett-Cotta. 105-120.

Stone, Philip J. (1980): A Perspective on Social Science Data Management. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 444-454.

Teddlie, C./Tashakkori, A. (1998): Mixed Methodology. Thousand Oaks et al.: Sage.

Thaller, Manfred (1986a):Vorschlag für einen internationalen Workshop über internationale Quellenbanken. In: Thaller, Manfred (Ed.): Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 20. St. Katharinen: Scripta Mercaturae . 9-30.

Thaller, Manfred (1986b): A Draft Proposal for a Standard for the Coding of Machine Readable Sources. In: Historical Social Research 11 (4). 3-46.

Thaller, Manfred (Ed.) (1986c): Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung. Historisch-Sozialwissenschaftliche Forschungen Volume 20. St. Katharinen: Scripta Mercaturae.

Thaller, Manfred (1987): Clio – Ein datenbankorientiertes System für die historischen Wissenschaften: Fortschreibungsbericht. In: Historical Social Research 12 (1). 88-91.

Thaller, Manfred (1988): Data Bases v. Critical Editions. In: Historical Social Research 13 (3). 129-139.

Thaller, Manfred (1991): The Historical Workstation Project. In: Historical Social Research 16 (4). 51-61.

Thome, Helmut (1989/1990): Grundkurs Statistik für Historiker. 2 Volumes. Historical Social Research Supplement 2 & 3.

Tidswell, David (1989): Scottish Internal Migration, 1812-1820: Interfacing Database and Computer Graphics Packages. In: Historical Social Research 14 (3). 42-47.

Trainor, Rick H. (1987): Introducing Microcomputers into History Teaching and Research: the Dish Project. In: Historical Social Research 12 (1). 72-75.

Treffeisen, Jürgen (2004): Archivische Überlieferungsbildung bei konventionellen Unterlagen im deutschsprachigen Raum. Eine Auswahlbibliographie. In: Historical Social Research 29 (4). 227-265.

Trugenberger, Volker (1992): Archivalien-Erschließung mit EDV in der staatlichen Archivverwaltung Baden-Württemberg: das Beispiel Reichskammergerichtsakten. In: Historical Social Research 17 (3). 136-141.

Tuchman, Gaye (1994): Historical Social Science. Methodologies, Methods and Meanings. In: Denzin, Norman K./Lincoln, Yvonna S. (Eds.) (1994): Handbook of Qualitative Research. Thousand Oaks/London/New Delhi: Sage. 306-323.

Van den Nieuwenhof, Patrick (2003): Archivilization of Sciences Archives. New Techniques Making Science Archives Understandable. In: Historical Social Research 28 (4). 242-255.

Vinovskis, Maris A. (1980): Problems and Opportunities in the Use of Individual and Aggregate Level Census Data. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 53-70.

Volkens, Andrea/Bara, Judith/Budge, Ian (2009): Data Quality and Content Analysis. The Case of the Comparative Manfesto Project. In: Historical Social Research 34 (1). 234-252.

Waldow, Florian (2001): The Suggestive Power of Numbers. Some Remarks on the Problem of the Accuracy of Quantitative Indicators in Comparative Historical Research. In: Historical Social Research 26 (4). 125-140.

Weber, Max (1906-1922): Die Wirtschaftsethik der Weltreligionen. Reprinted in: Gesammelte Aufsätze zur Religionssoziologie. 3 Volumes. UTB: Stuttgart.

Werner, Thomas (1991): Transforming Machine Readable Sources. In: HSR 16 (4). 62-73.

Wettengel, Michael (1995): Archivierung maschinenlesbarer Datenbestände im Bundesarchiv. In: Historical Social Research 20. (4). 123-126.

Winchester, Ian (1980): Priorities for Record Linkage. A Theoretical and Practical Checklist. In: Clubb, Jerome M./Scheuch, Erwin K. (Eds.): Historical Social Research. Historisch-Sozialwissenschaftliche Forschungen Volume 6. Stuttgart: Klett-Cotta. 414-430.

Witzel, Andreas/Medjedović, Irena/Kretzner, Susanne (Eds.) (2008): Secondary Analysis of Qualitative Data. Historical Social Research 33 (3). 7-214.