Data integration and consolidation of administrative data from various sources: the case of Germans' employment histories

Köhler, Markus; Thomsen, Ulrich

Data Integration and Consolidation of Administrative Data From Various Sources. The Case of Germans' Employment Histories

Markus Köhler & Ulrich Thomsen [*]

Abstract: *»Datenintegration und Datenkonsolidierung von administrativen Daten aus unterschiedlichen Quellen am Beispiel von Daten zum deutschen Arbeitsmarkt«*. This article introduces the data integration and consolidation process of the research data base of the Institute for Employment Research. The data are process generated data and stem from various, autonomous administrative processes. This fact implies that there are manifold inconsistencies between the data from the different data sources. This opens up the methodological problem of a successful consolidation of inconsistencies. Two contrarian strategies to handle this methodological problem are discussed and the solution in the IAB-data base is presented.

Keywords: Longitudinal Analysis, Process-Generated Data, Social Bookkeeping Data, Public Administrational Data, Data Management Record Linkage, Data Fusion, Labour Market Data.

1. Introduction

Process-produced micro data today play an important role in research in social science (cf. Jacobebbinghaus 2007). They are suitable for a broad range of research objects like research on organizations, administrative processes, employment biographies or special sub-population in a society (cf. Scheuch 1977: 22 et sqq.; Bick, Mann, Müller 1984; Schmähl, Fachinger 1994: 179 et sqq.; Bick, Müller 2002: 230; Wallgren 2007: 4). Process produced data are defined as data that are generated by diverse (administrative) processes and represent a documentation of different actions (cf. Scheuch 1977: 25).

The linkage between the administrative process and the data it generates implies that process-produced data necessarily are limited to a specific purpose. In order to overcome these natural limitations in the process generated data a combination of data from different data sources (and administrative processes) is necessary. At a first glance this data integration and consolidation process might look similar and typical for the development of research datasets. But

[*] Address all communications to: Markus Köhler, Institute for Employment Research, IT Services and Information Management, Regensburger Str. 104, 90478 Nürnberg, Germany; e-mail: markus.koehler@iab.de. Ulrich Thomsen, Institute for Employment Research, IT Services and Information Management, Regensburger Str. 104, 90478 Nürnberg, Germany; e-mail: ulrich.thomsen@iab.de.

using and preparing process-produced data stemming from large administrative processes poses some specific problems that are more complex those in data from smaller data sources.

So this article deals with data integration and data consolidation in general, but also depicts specific problems of the integration of large process generated datasets. In the following we want to present as an example the research data genesis of the Institute for Employment Research (IAB) data base that transforms and integrates data originating on the operational level of the appropriate administrative process of the Federal Employment Agency (BA) to datasets that are well fitted for research.

The above mentioned problems are not only technical but also methodological problems. The integration of data stemming in case of the IAB data from at least four different fields of administrative processes, two different institutions and, coming with that, minimum four different operational IT-systems[1] requires a dealing with inconsistencies in the data. So the methodological problem of how to deal with not integrated and inconsistent process generated data when creating a integrated research data base will be discussed and two diverging solutions will be presented.

All this information high lightens the background of the IAB data base and provides necessary knowledge for the usage of process produced data in general and especially the IAB research data, contributing to a better understanding of the data and data production process as it is proposed by Bick and Müller (1984: 145; cf. Wallgren 2007).

Before describing the data generation and the resulting problems the IAB research data base will be presented to ensure a better understanding of the following problems. This integrated research data base can serve as an example for integrated process data bases of various sources and fields.

2. Data Base of the Case Study: The IAB data

The IAB research data base consists in general of four different data sources and administrative processes covering the four main fields of the German labor market. These four fields are (a) employment, (b) job-search, (c) unemployment benefit recipience (unemployment insurance and basic income support) and (d) active labor market policies. The data are collected in different and autonomous operational IT-systems.

[1] In fact many more operational IT-systems were and currently are in use by the BA to administer and register process data, but we use in this article only the three main ones used for BA-based research data.

2.1 General Problems Arising When Integrating Data

In a first step the data of each data source are consolidated and for each data source a research data product is generated. In a second step these different data sources are integrated to one data set. This large and comprehensive research data base is called the Integrated Employment Biographies (IEB). In the last years many analyses of a broad range of aspects of the German labor market were based on the IEB. For more details on the IEB and the access to this data set, see Jacobebbinghaus 2007.

For a broader understanding of the presented IAB-data the four main data sources will be briefly introduced. Job-seeking data, unemployment benefit recipient data and data on measures of active labor market policy programs stem from the Federal Employment Agency itself and its operational IT-systems. These data are produced in the appropriate administrative processes. They are registered on the operational level in the belonging IT-system and then stored in the central pool of BA-business data. This data pool provides the data which are transformed and integrated by the IAB in order to build up its research data base.

In this process a whole set of problems can arise that also can be found in many other data integration processes. Some important problems are discussed in the following:

- *Inconsistent information for one person:* It is important to notice that there is nearly no cross-validation of information between the different operational IT-systems during the registration process. Therefore inconsistent information for one person (e.g. differences in nationality) can be registered.
- *Identifier (ID) of a person:* A second dimension of the autonomy of the different IT-systems is related to the identifier (ID) of a person. Different IT-systems and administrative processes do have different identifiers so that one person may have two or more identifiers.
- *Data Structure:* In principle the data structure in all data sources is similar. All the data are spell data. But coverage and composition of these spell data differs between the data sources. For example a spell in the data coming from the social insurance system maximal covers one year, whereas spells stemming from the unemployment benefit recipient system might be much longer.
- *Time-Span Covered:* The maximum time span covered in the data depends on the underlying administrative process. Some data sources already exist for decades, reaching back to the 1970s. This results in manifold problems that can cause breaches in the long time-series, like changes in legislation or the IT-systems, etc.
- *Different Target Populations:* The administrative processes providing the data are focussed on their special purpose and therefore target a specific group of persons. This implies that only information relevant for that purpose are registered and even further, that the meaning of the information is col-

lected purpose-dependent. Therefore in the process generated data seemingly similar attributes might have different, process dependent values and meanings.

2.2 Data Structure

The three BA-related data sources use the BA-client-number (KNR) as the main identifier for a person. The data consist of spell data, which means individual, daily and historical information, that show the different states of information over time. Included are client characteristics and different administrative purpose dependent information (e.g. target job in job-seeking information, amount of unemployment benefits per time period in unemployment benefit recipience, etc.).

In the case of *job-seeking data* this means individual, daily and historical information covering the time period since 2000 up to today, the different states of job-search and information about personal attributes like sex, date of birth, reasons for job-search, target job, etc.

Data stemming from the administration of *active labor market policy programs* also consist of individual, daily and historical information covering the time period since 2000 up to today. Information about personal attributes like sex, date of birth and information about the participation in the measure of active labor market policy is provided (e.g. duration of participation, outcome, etc.). Nearly all existing measures of active labor market policy since 1999 are covered in the data.

A significant longer time series can be found in the *unemployment benefit recipient data*. Here the spell data go back to 1975, so that more than 30 years are covered. The data show information about the unemployment insurance system. This means that only persons that were employed for a certain amount of time before becoming unemployed can be found in the data. The data show on a daily basis client characteristics, information about benefit recipience (e.g. amount of transfer per period), reasons for exiting the unemployment benefit recipient system and much more.

The youngest data source is the *basic income support system data*, which was founded by the introduction of the Social Code 2 (SGB II) in 2005. Since then there is a second kind of benefit recipience information coming from the administrative procedures of the basic income support system. The information found here is similar to the information in the unemployment benefit recipient data, but does have another target population and a different benefit logic – this system is a means-tested benefit system and in most cases it is conjointly administered by the BA and the municipalities.

Beside these data coming more or less from BA related processes, there is another large data source stemming from a completely different institution. The IAB hosts German employment data coming from the *social insurance system*.

So this data shows social insured employment in Germany, but no self-employment and no civil servants. It is important to notice that these data are produced by completely different institutions, institutions that are completely unrelated to the BA (see also Meyer, Mika and Thorvaldsen, in this Special Issue). The pension and health insurance institutions therefore use their own person identifier, the social security number. Analog to the unemployment benefit recipient data, the employment data start in 1975 and today end in 2007. The data consist of spell data, providing daily, historical and individual information about employment. Available information are personal characteristics, employment times, remuneration, occupation, etc.

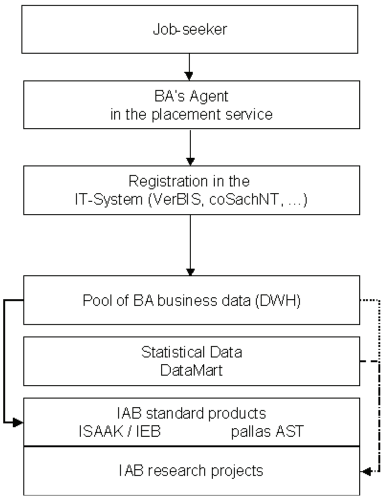
3. Data Generation and Institutional Bias

After the brief introduction of the most important parts of the IAB research data base, a closer look to the data generation process is necessary. This is essential, because these data, used for social research, are produced in a "foreign" – non scientific – context. The research data in this understanding represent a secondary usage of data that are collected for administrative purposes. Therefore administrative data being used as a research data source represent a data source of unobtrusive measures (cf. Best 2003: 112). To understand the research potentials and risks in the data background knowledge about the data generation process, situation and setting is required. Also important is knowledge about the administrative processes in which the data are collected (cf. Bick, Müller 1984: 123; cf. Wallgren 2007: 180). In order to give here a short impression of such background knowledge the genesis of the research data from the operation registration to the finished research data set will be outlined. Figure 1 shows the way from operational level to the research data using job-seeking data as example.

In the first step of the process the job-seeker communicates with the BA-agent in the placement service. The job-seeker is obliged to give information to the agent and these are only information relevant for the administrative process of job-placement. Some information are registered in the face-to-face situation of agent and client, so that the agent simultaneously has to communicate with the client and register data in the operational IT-system - and all that possibly under time pressure. This stressful situation has some implications on data quality. There might be some errors in the registered information and information more important for the placement process (e.g. occupational information) will be registered and checked priorly.[2]

[2] This is a general factor in administrative data. Process produced research data always are determined by the specific administrative system upon there are based (cf. Wallgren 2007: 176).

Figure 1: Model of Data Flow



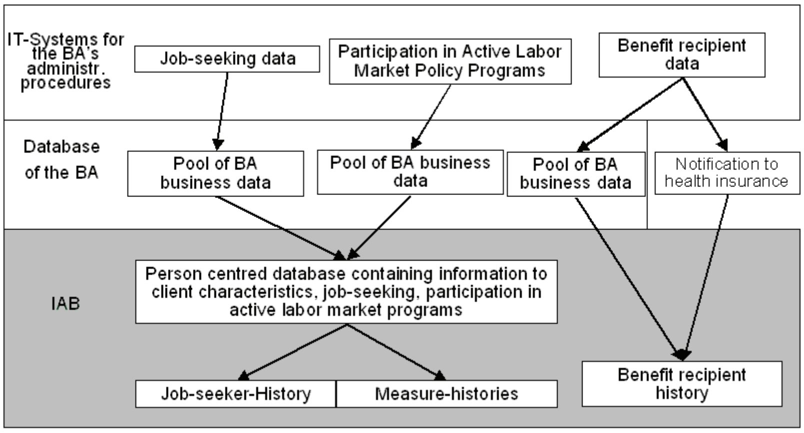
3.1 Lack of Cross-Consolidation

Another important point to notice is, that the IT-system is directly linked to the appropriate administrative process. So information concerning job-search is registered in the job-search IT-system (called VerBIS), information for the benefit recipience process in the benefit recipient IT-system. As mentioned above, there is nearly no validation of information by cross-checks between the different autonomous IT-systems.

After the registration and storage in the operational IT-systems some information is transferred to the central pool of BA business data. The pool centrally stores important data from the operational IT-systems and therefore includes data from all operational IT-systems of the BA plus some more selected data like employment data. It is important to notice, that the data are stored centrally, but are not integrated.

This data pool provides the base data used for creating the IAB data products. The data integration and consolidation work in order to build up the research data base is done by the IAB itself. The following graphic shows this explicitly:

Figure 2: Model of Data Sources


IT-Systems for the BA's administr. procedures
Job-seeking data
Participation in Active Labor Market Policy Programs
Benefit recipient data
Database of the BA
Pool of BA business data
Pool of BA business data
Pool of BA business data
Notification to health insurance
IAB
Person centred database containing information to client characteristics, job-seeking, participation in active labor market programs
Job-seeker-History
Measure-histories
Benefit recipient history

In this picture the pillar principle in the BA-data is made visible. The different administrative processes have appropriate operational IT-systems that work more or less autonomous. In the database of the BA the data are centrally stored but not integrated. This step is done by the IAB when creating an integrated person centered database. Some more steps in the IAB data production process are not illustrated above, but it should be clear that our methodological problem of dealing with inconsistent information has be solved before an integrated research data base can be built up.

In the next part three typical data problems will be presented and some examples for solutions are given:

1) The problem of the identification and integration of a person in the data will be illustrated.
2) Possibilities of consolidation of inconsistent information will be presented.
3) Some remarks about the problems resulting from long time series data will be made.

4. Integration and Consolidation of Data on a Single Person

As shown above the IAB integrates data from different sources. With that integration comes the need for consolidating manifold inconsistencies in the data. For example a person might have different nationalities at the same point of time in data source A and data source B.

A precondition for the dealing with inconsistent information is the identification of a comprehensive person. Only if one knows that person A in data source A is the same person as person B in data source B a consolidation of information between both data sources can be done.

So the first thing to be done is that person identification. Now why should that be a problem? As already mentioned different IT-systems use different person identifiers. We have at least two identifiers per person – the BA-client number (KNR) and the social security number (VSNR). In both cases it is possible that one person has more than one of those identifiers.

1) *BA-Client Number (Kundennummer, KNR):* Until 2003 the BA-client number was decentrally assigned by the different BA-agencies. Each agency of the BA (more than 270 in Germany) assigned its own number after a person entered the administrative process. That means that after changing the agency because of moving to another city a new number was assigned. A new number also was assigned if a client after a longer period of absence could not remember his old number and told the agent that he didn't have a BA-client number. Another factor is the above mentioned stressfull registration procedure. In this situation an already assigned BA-client number might be registered incorrect. All these factors lead to the fact that a person might have more than one BA-client number, either historically or even at the same point of time. Another identification problem can be found in the historical BA-data. Before 2005 more identifiers existed along with the BA-client number. An example later on used is the BA-registration number (EGNR).
2) *Social Security Number (Versicherungsnummer, VSNR):* The second major person identifier comes with the employment data. The social security number is assigned by the pension insurance. For each person that enters an employment that is covered by the social insurance system the pension insurance assigns a stable identifier that should not change over the life course. But here it is possible too that a person has more than one number.

All these problems result in the need for a stable, consistent and comprehensive person identifier in the integrated IAB data. There are two possible strategies to create that identifier:

1) Known relations between the different identifiers in the BA data can be used. By combining the already known related identifiers one person can be identified as being one person in the different data sources.
2) Known personal attributes like name, first name, date of birth and sex are matched. If that matching brings a good degree of accordance it is plausible that a person in two data sources actually is the same person.

In the case of the IAB data both possibilities are used and will be briefly outlined here.

4.1 Identifying and Checking Connections

In a first step all the given identifiers in the data are systematically checked for connections. This is possible because in the different IT-systems along with the primary ID of that operational IT-system other IDs are registered as secondary information. An example for this are the job-seeking data. Here the primary identifier is the BA-client number (KNR), but during the administrative process the social security number (VSNR) is registered as well as a secondary attribute. This results in relations between the different identifiers in different data sources. By systematically checking these connected information a person and its different IDs can be identified.

The second step for identifying a comprehensive person is matching four personal attributes and look for the accordance in that match. The second step is possible because of a (redundant) storage of this information in different IT-systems. The four already described personal attributes can be found in the different data sources. And this step is necessary for avoiding wrong mergers of persons.

The following example with fictive data in two tables makes this identification process transparent. The fictive example consists of data of two persons who are married. The data used are stored in two central table, the BA-Client-History-Table and the BA-Client-Table. The identification process uses combined information from both table. The example consists of three steps: First related identifiers in table 1 are combined, second the related identifiers in table 1 are combined with identifiers in table 2 and third the proposed combination of identifiers are validated by matching personal attributes. In the end the two persons are comprehensively identified.

The first table used in this example is the BA-Client-History:

Table 1: Fictive BA-Client-History

	BA-Client-History				
KNR	VSNR	SEX	Date of Birth	Name	First name
12	13111080S611	M	11.10.80	Maier	Hans
14	13111080S611	M	11.10.80	Maier	Hans
14	52111080S613	M	11.10.80	Maier	Hans
15	13111080S611	W	01.12.82	Maier	Ingrid
17	13111080S611	M	11.10.80	Maier	Hans

The second table to be considered is the BA-Client-Table:

Table 2: Fictive BA-Client

BA-Client	
KNR	EGNR
12	123
12	124
13	124
15	125

The identification process goes as follows:

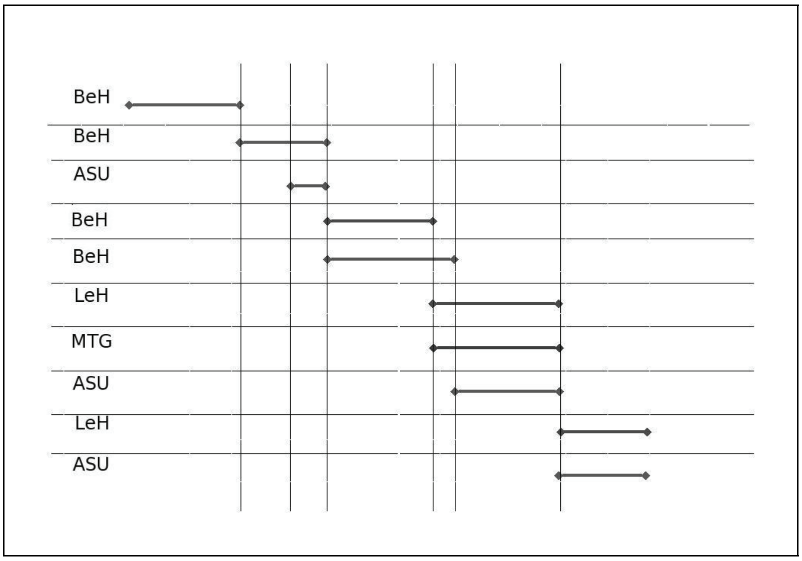
1) In the first step the given relations between the IDs are used. Using the BA-Client-History, VSNRs 13111080S611 and 52111080S613 are allocated to the same person via KNR *14* in BA-Client-History. KNR *12*, *14* and *17* are also assigned to that person because of having the same VSNR.
2) Taking into consideration the EGNR in the BA-Client-Table KNR *13* can be related to KNR *12* via the identical EGNR *124*. Therefore KNR *13* refers to this person too.
3) Going one step further KNR *15* could also be related to this person because of the identical VSNR (see BA-Client-History). At this point the second i-dentification step gets important. KNR *15* has a different sex, date of birth and first name. When matching these four attributes no accordance can be found. Therefore KNR *15* is identified as being a second person. Considering the BA-Client-Table EGNR *125* is allocated to that second person.

So in the end, linking the different related IDs would indicate one identical person in all the given data, but matching the already mentioned four central personal attributes result in an inconclusive match. Therefore we assume in this case the related IDs would wrongly be linked. This results in the identification of two separate persons. Person *1* has four different KNRs, two different VSNRs and two EGNRs. Person *2* has one KNR, one VSNR and one EGNR. This reduces the given five KNRs, two VSNRs and three EGNRs to two comprehensively identified and checked person-IDs.

4.2 Identifying and Handling Data Integration Inconsistencies

Now that we have an integrated and comprehensive person, the integration of the data can begin. In this step of the data production process inconsistencies in the data become visible. As mentioned above, there is nearly no cross-checking of the data during the data registration process. So in the administrative process only few of the registered information in the different IT-systems are validated by comparing information between the different operational IT-systems. This can lead to inconsistencies. For example one person can be registered as being employed and unemployed at the same point of time, because these two status information are raised by two different administrative institutions (pension insurance and BA). But this also can happen in the BA data, when different administrative processes and IT-systems are used. The following figure, showing visualized spell data along the time axis from left to right, illustrates this fact:

Figure 3: Spell data



BeH: Time of employment | ASU: Time of job-search | LeH: Time of unemployment benefit recipience | MTG: Time of participation in active labor market policy programs

It can be seen that different data sources and states do have overlapping time periods. On a first glance it is not clear which overlaps are plausible and which are implausible. Under some premises overlapping times of employment and job-search are allowed, but under other circumstances they are not. So legal background knowledge is required to decide which overlapping spells should be consolidated and which have to remain overlapping. A consolidation rule would have to account for all possible (and impossible) overlapping times and inconsistent states and information. A good example for those rules would be the consolidation of employment times and times of unemployment benefit recipience (SGB III). In general parallel times in this case are forbidden. But if the employment does not exceed a certain amount of working hours per week, parallel times of (small part time) employment and benefit recipience is allowed. On the other hand full time employment parallel to benefit recipience is not allowed under any circumstances. Here the exact knowledge of the legal background is required to decide about the correct consolidation.[3]

[3] It has to be noticed that such a combination of forbidden states might also be an indication of benefit abuse. In any case, if data error or benefit abuse, only an exact knowledge of the background information together with a sound analysis of the data itself allow an adequate interpretation and consolidation of the data.

All these sketched problems illustrate that a consolidation of data from two different data sources perhaps might be possible, but developing consolidation rules for more than two or three different data sources is very fast getting too complex to do.

4.3 Degree of Consolidation

Beside the complexity of the necessary consolidation rules a second fact has to be taken into account. Consolidating data always means overwriting and therefore destruction of information. This means that there are two diverging strategies concerning the consolidation of inconsistent data: Little consolidation versus much consolidation. The strategies have direct implications for the research that can be conducted with generated data base:

1) *Little consolidation*, coming with a smaller loss of information but many inconsistencies, generates more individual research options and more expenses for data preparation for the researcher. Following from this is in a broader variety of research results, but also perhaps disparate and incomparable findings.
2) *Much consolidation* overwrites information and this information is lost for the researcher. A stronger consolidation results in structured data, thereby also structuring the research options and minimizing the expenses for the researcher. An error in the consolidation rules inevitably leads to misinterpretations of the data, therefore possibly producing false research results.

One more problem in integrated administrative data will be presented here. This problem is directly related to the result of a strong consolidation strategy. Much consolidation leads to overwritten information. But in administrative data a seemingly similar attribute can have more than one process dependent meaning resulting in different ranges of values. Training is good example for this problem. Information about the level of training can be found in diverse data sources. But the exact meaning of the attribute "training" changes for each data source because it is directly related to the appropriate administrative process. In case of employment data the information in that attribute are related to employment, therefore more oriented on the level of education of the person. In the case of job-seeking data the seemingly same attribute has a different range of values and these values are skill-level oriented. That is the case because the administrative purpose is job-search where the skill-level is the important information. In such cases often it is impossible to correctly transcode the different information, so the process related information has to be kept in the data.

4.4 Handling Long Time-Series

The fourth subject to be discussed here is the time dimension of the data. As mentioned in chapter 2 the available time series of employment and benefit recipient data starts in 1975. This means legal configurations could have changed and in fact did change over the last 30 years, as well as the administrative practices, the IT-systems, etc. Even formerly forbidden states in the data today could be allowed or vice versa.

Such long time series in process generated data provide special challenges. The mentioned changes in the legislation, changes in the administrative processes and practices and changes in the operational IT-systems all result in breaches in the time series of the historical data. Moreover the classification of occupation and economic sector, the range of values in attributes and the regional information can and will change over time. Another big challenge poses the old documentation, that is rudimental at best. Getting background information regarding IT-systems or administrative practices from documentation from the 1970s or the 1980s is nearly impossible. The experts from that times are retired and therefore the knowledge is inaccessible.

So, concerning historical data, there is no golden or general rule for handling the data.

In the case of changes in legislation or changes in the IT-systems the smoothing of the breaches depends on the dimension of the change. A breach in the time series is inevitable if an exact mapping of information from the older data structure to the newer one is impossible. This is due to the fact that without an exact mapping the transformation of older information into the newer data structure would mean a massive and back dated overwriting of information. It also would mean to inscribe data from the newer situation into a different historical situation. Therefore the changes in the data only can be documented if an exact mapping is not possible. This leaves the researcher free to decide how he interprets and deals with the breaches.

The same fact holds true for changing ranges of values in attributes. A good example for that is the classification of economic sectors. This classification changed three times in the last fifteen years. In two cases an exact one to one mapping of old classification to new classification was not possible. In the last change in about 30% of the old classifications the mapping to the new classifications is inconclusive. This implies that there is no possibility to consolidate the data. The solution here can only be describing and documenting the changes. A researcher using the data can choose himself how to deal with this breach in the time series of the data.

Another strategy is used when regional information have to be integrated, consolidated and updated. Regional information in many cases can be mapped to the newest status. This is possible because for any case of change in the dimension of the municipalities, there exists an exact mapping of old regional information to new regional information. An example for this would be the reformation of the size of districts in eastern Germany in the mid-1990s or in Saxony in 2008. In each case it is clear which part of district went to which new district or which districts were merged. This administrative information is available and hence can be used to consolidate the regional information when producing the integrated IAB data base.

Summing up the remarks concerning historical data, it can be said that in general a smoothing of the breaches in the long time series in the data is not possible. Therefore the main focus lies on the exact documentation of known facts and eventually of background knowledge.

5. Conclusion

To sum up the general guidelines, applicable for data errors stemming from the data registration process as well as problems in time-series, the following facts are consolidated:

- only valid values of an attribute are used
- only valid dates are used (e.g. start date < end date)
- filters and default values in the operational IT-systems are documented as far as they are known
- some central personal attributes like sex and date of birth are consolidated on the basis of the identification of a comprehensive person
- background knowledge and important legal or procedural knowledge is documented

This brings us to the answer of the general methodological questions asked in the introduction. There is no golden rule how to deal with inconsistencies in the data, because each problem solving strategy has specific implications and consequences. And these are general ones. An adequate answer to the question of data integration und consolidation can only be an individual one and depends on the specific project. But the consequences of each strategy always have to be kept in mind.

Our philosophy emphasizes weaker consolidation rules favoring documentation instead of hard data consolidation. This leaves more degrees of freedom to the researcher, enabling him to make the necessary decisions in data cleansing by himself and fitting to his priorities. The IAB research data base therefore is backed up by detailed and extensive documentation – a documentation that is essential to know when working with this research data base. And this strategy is applicable for other research data too.

The IAB data integration and consolidation processes result in standardized research data products, covering the complete targeted group with nearly no item- or unit-nonresponse, gaps in retrospection, selective retention or recall errors. On the other hand the data quality especially of data from the 1970s or 1980s is hard to grasp and, being administrative data, the data lack information on household, attitudes, preferences, etc.

The philosophy of data preparation in this data implies that only in clearly defined cases consolidation is done. The emphasis lies on documentation of the facts and background information. Therefore a good knowledge of the data generation process, its problems and framework is required and essential for using the full potential of the data.

References

Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (1984): Massenakten als Datenbasis der empirischen Sozialforschung. Methodische Voraussetzungen und institutionelle Erfordernisse. In: Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Hrsg.) (1984): Sozialforschung und Verwaltungsdaten. Stuttgart: Klett-Cotta. 9-18.

Bick, Wolfgang/Müller, Paul J. (1984): Sozialwissenschaftliche Datenkunde für prozeß-produzierte Daten: Entstehungsbedingungen und Indikatorqualität. In: Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Eds.) (1984): Sozialforschung und Verwaltungsdaten. Stuttgart: Klett-Cotta. 123-159.

Bick, Wolfgang/Mann, Reinhard/Müller, Paul J. (Eds.) (1984): Sozialforschung und Verwaltungsdaten. Stuttgart: Klett-Cotta.

Bick, Wolfgang/Müller, Paul J. (2002): Focus im Rückblick: Massenakten als Datenbasis der empirischen Sozialforschung. Methodische Voraussetzungen und institutionelle Erfordernisse. In: Historical Social Research 27 (2/3). 227-252.

Hauser, Richard/Ott, Notburga/Wagner, Gert (Hrsg.) (1994): Mikroanalytische Grundlagen der Gesellschaftspolitik: Ergebnisse aus dem gleichnamigen Sonderforschungsbereich an den Universitäten Frankfurt und Mannheim / Deutsche Forschungsgemeinschaft. Berlin: Akademie Verlag.

Jacobebbinghaus, Peter/Seth, Stefan (2007): The German Integrated Employment Biographies Sample IEBS. In: Schmollers Jahrbuch 127. 335-342.

Müller, Paul J. (Ed.) (1977): Die Analyse prozeß-produzierter Daten. Stuttgart: Klett-Cotta.

Scheuch, Erwin K. (1977) „Die wechselnde Datenbasis der Soziologie - Zur Interaktion zwischen Theorie und Empirie", in: Müller, Paul J. (Ed.) (1977): Die Analyse prozeß-produzierter Daten. Stuttgart: Klett-Cotta. 5-41.

Schmähl, Winfried/Fachinger, Uwe (1994): Prozeßproduzierte Daten als Grundlage für Sozial- und verteilungspolitische Analysen - Erfahrungen mit den Daten der Rentenversicherungsträger für Längsschnittanalysen. In: Hauser, Richard/Ott, Notburga/Wagner, Gert (Hrsg.) (1994): Mikroanalytische Grundlagen der Gesellschaftspolitik: Ergebnisse aus dem gleichnamigen Sonderforschungsbereich an den Universitäten Frankfurt und Mannheim / Deutsche Forschungsgemeinschaft. Berlin: Akademie Verlag. 179-200.

Wallgren, Anders/Wallgren, Britt (2007): Register-based Statistics. Administrative Data for Statistical Purposes. Chichester: John Wiley & Sons, Ltd.