German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2018-02 | APRIL 27, 2018

Record linkage of GDR's "Data Fund of Societal Work Power" with administrative labour market biography data of the German Federal Employment Agency

Manfred Antoni

Contents

Abstract

This working paper describes the record linkage of the ”Data Fund of Societal Work Power” of the former German Democratic Republic (GDR) with administrative labour market biography data of the German Federal Employment Agency. The linkage success rate was 77%. Due to legal restrictions, the resulting linked data are not yet available to the general scientific community.

Keywords: administrative data, German Democratic Republic, Germany, employment data, record linkage

Acknowledgements: I thank Dana Müller and Hannah Liepmann for valuable discussions during the whole linkage process and Robert Jentzsch and Silvina Copestake for the extract of linkage identifiers from the Data Warehouse of the Federal Employment Agency.

1 Introduction

This working paper describes the record linkage of the "Data Fund of Societal Work Power" (henceforth abbreviated as GAV from its German name *Datenspeicher Gesellschaftliches Arbeitsvermögen*) of the former German Democratic Republic (GDR) with administrative labour market biography data of the German Federal Employment Agency (BA). The linkage was done for the project "Labor Market Trajectories of East Germans around Reunification", a joint project of the Research Data Centre (FDZ) of the BA at the Institute for Employment Research (IAB) and the Humboldt-University Berlin. See Liepmann and Müller (2018) for more details on the project as well as on the GAV. As the GAV data were provided by the Federal Archive of Germany specifically for the linkage and analyses within this project, the linked data can not yet be accessed by the general scientific community.

Section 2 describes the two data sources and the non-unique linkage identifiers they contained. Section 3 provides details on the cleaning and standardising of these identifiers, before Section 4 describes the comparison and classification steps within the linkage process. Section 4 discusses the linkage results, and Section 5 provides a short summary and some concluding remarks.

As this paper focuses on the specific challenges and decisions within the project at hand, I do not attempt to provide a comprehensive overview of linkage techniques. For an in-depth coverage of the methods used here see, for instance, Christen (2012) or Herzog et al. (2007). Antoni and Schnell (2017) provide a general overview of the record linkage methods developed and applied by the German Record Linkage Center (GRLC, see also http://record-linkage.de), whereas Schild and Antoni (2014) provide more details on linkage applications involving the administrative data of the BA.

2 Data Sources

2.1 Data Fund of Societal Work Power (GAV)

The GAV was provided by the Federal Archive of Germany. While the administrative data described in Section 2.2 are longitudinal in nature, the GAV data extract available for this linkage contains cross-sectional employment data on slightly 7 million people living and working in the GDR at the reference date December 3, 1989. This data extract covers about 72% of the GDR's labour force at the end of 1989. Liepmann and Müller (2018) provide an overview of the origin and content of the GAV data. While they also discuss elements of the data that are suitable for substantive analyses, I will focus on features of the data that are relevant for the linkage process.

After some initial data preparation by Liepmann and Müller (2018, see pp. 10-11) to deal with duplicate entries, I was able to use the following linkage identifiers of almost 7 million people covered in the GAV extract:

- Given and family name (in one common field)

- Day, month and year of birth
- Sex

The quality of the variables on the date of birth and the sex of persons covered by the data extract was sufficiently good to facilitate a successful linkage. However, the given and the family name were stored in one common field, and the ordering of and the delimiters between different parts of the name was not entirely consistent within the data extract. The solutions chosen to deal with these issues are described in Sections 3 and 4.

Some identifiers that are commonly also used when linking data on persons or households were missing in the data extract at hand. Most importantly, any information on the residential address of the observational units at the reference date is missing from the GAV extract. While this limitation did provide some challenges for the linkage process, the solutions described in the upcoming sections made it possible to compensate for this shortcoming to a certain degree.

2.2 Administrative data of the German Federal Employment Agency

The linkage identifiers used to find people within the administrative data of the Federal Employment Agency (BA) were drawn from the Data Warehouse of the BA by IAB's department *Data and IT-Management*. These data stem from the following sources:

- mandatory social security notifications by employers about any of their employees subject to social security contributions (i.e., not including self-employed and civil servants)
- internal processes of the BA regarding
    - benefit recipients according to Social Code Books II and III
    - registered job-seekers
    - participants in active labour market policy measures

Data entries from any of these sources that relate to the same person are already integrated by the statistics department of the BA, where every person represented in the data also receives a unique and time-consistent person identification number. Note that this unique pseudonym is only valid within the data of the BA and cannot be used to identify persons in any external data source. Again, Liepmann and Müller (2018) provide additional details on the data available at the BA.

The following linkage identifiers were contained in the extract from the BA's administrative data:

- Given and family name (in separate fields)
- Birth name (in a separate field, only available in 3.4 percent of the records)

- Day, month and year of birth
- Sex

It would have been possible to additionally extract the residential address of the persons covered here. However, as the GAV did not contain any information on a person's residential address at the reference date in 1989, and because there was substantial mobility among former citizens of the GDR following the German reunification, this information would not have been useful in the linkage application at hand.

To uphold the principle of data economy, to make the linkage process more efficient and to avoid false-positive linkage results, I did introduce some restrictions regarding the extract from the BA's Data Warehouse. The extract therefore only contained linkage identifiers of people who

- were born between 1929 and 1976, as people of earlier birth cohorts would probably have retired instead of showing up in the German administrative data, whereas younger birth cohorts would have been 13 at the end of 1989, which makes it very unlikely that they are included in the GAV extract;
- did not show up in any administrative records of the BA prior to the year 1990, as that would have meant that they had already been living in West Germany prior to the reunification;
- did have at least one record in the data of the BA between 1990 and 1996, either in East or in West Germany.

The combination of these criteria lead to a selection of people into the data extract that had a relatively high likelihood of having lived in the GDR until the reunification in 1990. Apart from considerations of data economy or data handling efficiency, such a strict a-priori restriction of possible linkage candidates was necessary to compensate for the limited set of linkage identifiers and to avoid false-positive matches. Despite these restrictions, the data extract still contained about 21 million records of almost 16 million people.[1]

3 Preprocessing

Although both data sources originate from data generating processes that inherently include some quality checks, errors or inconsistencies during the collection process cannot be ruled out entirely. All of the linkage identifiers described above therefore have to be considered to be error-prone, and using them for a comparison without any cleaning in advance may lead to a suboptimal linkage result. The GRLC has developed a variety of preprocessing scripts

[1] Note that there was a strict limitation on who of the project members was allowed to access certain parts of the data sources mentioned above. While the project members and authors of Liepmann and Müller (2018) had the permission of the Federal Archive of Germany to access the entire GAV extract, only I was permitted to access the linkage identifiers in both the GAV extract and the administrative data of the BA.

in order to clean and standardise name and usually also address fields. These routines are described in more detail by Schild and Antoni (2014).

The basic steps to modify string variables in both data sources included the following:

- Replacing German umlauts and other non-ASCII characters with ASCII equivalents
- Changing all characters to uppercase
- Removing leading and trailing blanks
- Removing punctuation and special characters

However, as the GAV extract did only contain one common field including both given and family name, while the administrative data of the BA already provided separate name fields, the preprocessing of the name field(s) had to be done differently for the two datasets. For example, removing academic titles, generational assignations and titles of nobility from the data of the BA could be done relatively easily, as such pre- or suffixes did only occur at the beginning or the end of either of the name fields. In the GAV extract, however, such elements were placed inconsistently within the common name field, i.e. either at the beginning, in the middle or at the end. Other inconsistencies within the GAV extract made the cleaning of the name field and the parsing of its relevant elements in separate fields even more complex, e.g. a varying ordering of the first and last name or inconsistent or even missing delimiting characters.

There were additional issues that made cleaning steps specifically for the GAV extract necessary. First, a small share of records (about 16,000 in total, see Liepmann and Müller, 2018, pp. 11-12) contained a name field that was too short to plausibly contain a valid full name. In many of these cases that field did only contain one (truncated) part of the name. Given that the limited set of linkage identifiers, especially the lack of any residential information, would not have allowed to compensate for a missing valid name, we removed all record with common name fields with up to 4 letters.

Second, even name fields with more than 4 letters often did only contain one word (for about 500,000 persons, see Liepmann and Müller, 2018, pp. 11-12), i.e. a string of letters without any delimiters in it.[2] Such fields most often only contained the family name. For the majority of these cases the information was used to fill the separate family name field, while the first name field remained empty. For a small share of these cases we could parse a given name from a field without any delimiters by creating lists of very common female and male names from the separate name field available in the data of the BA. However, to avoid false replacements resulting from this procedure, these cases did undergo a clerical review, which in turn lead to a number of manual corrections.

One special feature of the data of the BA was a separate field for the birth name in case it was different from the registered family name. This field was preprocessed consistently to the

[2] Any existing punctuation in this field in the GAV extract had not been removed at this point of the process. The only exceptions were sequences of multiple delimiting characters, which were replaced by single commas.

family name field. However, the birth name field was only filled in 3.4 percent of the records, which probably means that there would have been more people with differing birth name and current family name, but that difference was not registered on many cases.

4 Comparison and classification

4.1 Brief description of methods

The cleaning and parsing of the linkage identifiers was followed by several steps of comparison of record pairs from the two data sources. This section provides a brief description of the methods applied in these comparisons.

Deterministic linkage: In a deterministic linkage step, all or a predefined number of identifiers have to agree exactly. If the required number of agreeing identifiers is fulfilled, a record pair is classified as a match.

Distance-based linkage: In real-world data, even after the preprocessing, a considerable share of records usually still has errors of some sort, e.g. typos, misspellings, use of abbreviations, or different orderings of name components. In a deterministic linkage setting, even the smallest deviations between identifiers could lead to a classification as a non-link, even though the record pair might in fact be a true match. Distance-based linkage deals with such deviations by computing the similarity of different identifier representations and by matching record pairs that exceed a certain similarity threshold. In the distance-based linkage steps described below, I use the bi-gram similarity[3]. I use the Merge ToolBox (MTB) software[4] to perform the distance-based linkage.

Blocking: Blocking limits the number of comparisons between record pairs necessary and thereby considerably reduces computing time. This is achieved by only comparing record pairs that have equal values on one or more blocking variable, e.g. on sex and/or on the birth year.

Array matching: When an identifier has more than one possible representation within one of the compared datsets, array matching can be applied to increase linkage success. In the project at hand, the BA data contain the family name and sometimes also the birth name for the same person, whereas the GAV only contains he family name. An array match compares all representations of that identifier in the first dateset with all representations of that identifier in the second dataset and matches the record pair with the highest similarity value of all these comparisons.

[3] Bi-grams are substrings of a string with the length 2, e.g., the bi-grams for the word MARY would be MA, AR and RY.

[4] The MTB is maintained by the GRLC and can be downloaded and used for free for academic purposes. See http://record-linkage.de or Schnell et al. (2004) for more details on the software.

4.2 Linkage results

Table 1 summarizes the linkage steps that were done consecutively, starting with the strictest criterion of an exact agreement on all available identifiers (step *Deterministic 1*). In this step, 68 percent of the 6978591 persons available in the GAV extract are already linked successfully.

Table 1: Summary of linkage steps

	N	share	description
	6978591	100.00%	Number of persons in GAV before linkage
Linkage steps			
Deterministic 1	4760376	68.21%	Exact agreement on given name, family name, sex, birth date (day, month, year)
Distance-based 1	149620	2.14%	Bi-gram similarity of given name and family name; exact agreement on day of birth date; blocking on month and year of birth date, sex
Distance-based 2	9662	0.14%	Bi-gram similarity of one common name field (family name and given name concatenated); exact agreement on day of birth date; blocking on month and year of birth date, sex
Deterministic 2a	26535	0.38%	Exact agreement on one common name field (family name and given name concatenated), sex, month and year of birth date
Deterministic 2b	24001	0.34%	Exact agreement on one common name field (birth name and given name concatenated), sex, month and year of birth date
Deterministic 3a	315682	4.52%	Exact agreement on family name, sex, birth date (day, month, year); combination must be unique; all compared fields need to be filled
Deterministic 3b	11516	0.17%	Exact agreement on birth name, sex, birth date (day, month, year); combination must be unique; all compared fields need to be filled
Deterministic 4	110425	1.58%	Exact agreement on given name, sex, birth date (day, month, year); combination must be unique; all compared fields need to be filled
Total	5407817	77.49%	

Source: Date Warehouse of the BA, GAV; own calculations.

Only persons not successfully linked in this first step are then compared in step *Distance-based 1*. In this step, bi-gram similarities are computed for the identifiers given name and family name, whereas only exact agreement is evaluated for the day of the birth date. To reduce computation time, blocking is used on the month and year of the birth date as well as on sex. Relative to the previous step, the contribution of this first distance-based step is rather small, as it only adds 2.1 percentage points to the overall linkage rate. Step *Distance-based 2*

is identical to the first distance-based step except for the fact that now the family name and the given name are concatenated into one common field in each of the datasets. This field is then compared via the bi-gram similarity. The contribution of this distance-based step is even smaller than that of the previous one, as it adds only .14 percentage points to the overall success rate.

Step *Deterministic 2a* combines the strictness of *Deterministic 1* with the common name field first used in step *Distance-based 2*, which replaces the separate fields for family name and given name. As this step is very similar to the very first one, it is not surprising that it only adds .38 percentage points to the linkage rate. Step *Deterministic 2b* only deviates from the previous step by replacing the family name in the concatenated common name field with the birth name. Again, with .34 percentage points the contribution of this step is rather low. As the birth name is only filled in 3.4 percent of the records in the BA data, this finding is not surprising.

Steps *Deterministic 3a* and *Deterministic 3b* both require exact agreement on sex and the full birth date, but they differ in the name field they use in the exact comparison (family name and birth name, respectively). Both steps have in common that the strictness seems to be reduced by not also requiring the given name to agree. However, record pairs fulfilling either of these steps' criteria also need to be unique in both of the datasets. Thus, if any combination of representations for the compared identifiers is found more than once even in one of the two datasets, all such record pairs are disregarded for the current step. An additional requirement is that all of the compared fields need to be filled. This requirement makes sure that no record pairs are accepted based on agreement on mutually missing identifiers. Together, these two steps add 4.69 percentage points to the linkage success rate.

Step *Deterministic 4* is identical to the previous two steps, except that instead of the family name or birth name, the exact comparison uses the given name. This final step adds 1.58 percentage points to the overall success rate.

5 Discussion

The overall linkage success rate was about 77 percent, which is lower than in previous linkage applications with data of the BA.[5] However, previous linkages could rely on sets of linkage identifiers that were far superior to that in the application described here. Moreover, these and other previous linkages did not have to deal with a considerable time-lag between the collection periods of the different datasets. In the project at hand, the likelihood that people included in the GAV extract have retired or otherwise withdrawn from the labour market before ever showing up in the data of the BA is considerably higher than in other linkage applications. Given that the available set of linkage identifiers was rather limited and that the administrative data of the BA already only cover roughly 85 percent of the German workforce, the linkage can be classified as rather successful.

[5] Antoni et al. (2018) and Antoni et al. (2017) report linkage success rates of 90.0 percent and 85.4 percent, respectively.

As of yet, the linked data described here cannot be made available to the general scientific community. The project members are currently working towards a solution to overcome the remaining legal restrictions and to facilitate future access to the linked data via the FDZ.

References

Antoni, M. and R. Schnell (2017). "The Past, Present and Future of the German Record Linkage Center (GRLC)". In: Journal of Economics and Statistics. Online first.

Antoni, M., S. Dummert, and S. Trenkle (2017). PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975-2015. FDZ-Datenreport 06/2017 (de).

Antoni, M., N. Bachbauer, J. Eberle, and B. Vicari (2018). NEPS-SC6-Erhebungsdaten verknüpft mit administrativen Daten des IAB (NEPS-SC6-ADIAB 7515). FDZ-Datenreport, 02/2018 (de).

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin: Springer.

Herzog, T. N., F. J. Scheuren, and W. E. Winkler (2007). Data quality and record linkage techniques. New York: Springer.

Liepmann, H. and D. Müller (2018). A proposed data set for analyzing the labor market trajectories of East Germans around reunification. FDZ-Methodenreport 03/2018 (en).

Schild, C.-J. and M. Antoni (2014). Linking Survey Data with Administrative Social Security Data - the Project "Interactions Between Capabilities in Work and Private Life". Tech. rep. German RLC Working Paper No. wp-grlc-2014-02.

Schnell, R., T. Bachteler, and S. Bender (2004). "A toolbox for record linkage". In: Austrian Journal of Statistics 33, pp. 125–133.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editors

Rainer Schnell, Manfred Antoni

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center was funded by the German Research Foundation (DFG).