German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2024-01 | FEBRUARY 21, 2024

Linking graduate data from the Nuremberg Institute of Technology and administrative labor market biography data from the Institute for Employment Research

Manfred Antoni | Timon K. Drewes | Hans-Dieter Gerner | Robert Jäckle | Stefan Schwarz

Linking graduate data from the Nuremberg Institute of Technology and administrative labor market biography data from the Institute for Employment Research

Manfred Antoni (Institute for Employment Research, IAB)
Timon K. Drewes (University of Erlangen–Nuremberg, FAU and Leibniz Institute for Educational Trajectories, LIfBi)
Hans-Dieter Gerner (Nuremberg Institute of Technology, THN and Competence Center for Social Innovations, Methods and Analyses, KoSIMA)
Robert Jäckle (THN and KoSIMA)
Stefan Schwarz (IAB)

German Record Linkage Center: Working Paper

Contents

Abstract

This paper documents the process of linking the records of all bachelor's and master's graduates of the Nuremberg Institute of Technology between 2010 and 2020 with administrative labor market biography data from the Institute for Employment Research (IAB). The success rate of the linkage was 98%, which is well above the average of previous linkage projects with IAB data. We only find negligible differences between characteristics of matched and unmatched graduates. Notably, linkage success is lower for foreign graduates than for German graduates. On average, the time to graduation is shorter for unmatched than for matched graduates.

Keywords: administrative data, employment data, Germany, higher education, record linkage

Acknowledgements: We thank Christian Teichert for his support and information he provided to bring this project forward and foster further progress in record linkage projects dealing with university and administrative data. We also thank Lisa Schmidtlein and Stephan Grießemer from IAB's department Data and IT-Management for providing the linkage identifiers from the Data Warehouse of the Federal Employment Agency. We gratefully acknowledge administrative and legal support from the Nuremberg Institute of Technology. Our special thanks go to Matthias Wichmann for providing the university data and to Raphael Brade for pre-processing the university data before it was transferred to the IAB.

1 Introduction

This working paper documents the record linkage of all bachelor's and master's graduates from the Nuremberg Institute of Technology (*Technische Hochschule Nürnberg Georg Simon Ohm*, THN) with administrative labor market biography data from the Institute for Employment Research (*Institut für Arbeitsmarkt- und Berufsforschung*, IAB) of the German Federal Employment Agency (*Bundesagentur für Arbeit*, BA). The linkage was a joint project of the THN with IAB.

So far, research that links students' university data with administrative labor market biography data is still scarce (see Kaul et al., 2016; Möller, 2017; Teichert, Liefner, and Otto, 2021; Teichert, Niebuhr, et al., 2020 for projects using German data and Comunian, Jewell, and Faggian, 2017 for a project using data from the UK). Our project aims to foster progress in this field, and this working paper documents the process of linking university records and administrative labor market biography data for further research in that field.

Our record linkage achieved a remarkable success rate of 98%, exceeding the average of previous linkage projects using IAB data. We find only minimal differences between the characteristics of matched and unmatched graduates. In particular, the linkage success rate was lower for graduates with foreign citizenship than for German graduates, possibly because there are some students in this group who may return to their home countries after graduation. In addition, unmatched graduates tended to have a shorter average time to degree than their matched counterparts.

The linked graduate-labor market data can be used to examine graduates' employment spells during and after their studies. In addition, the data enable us to analyse educational decisions and their determinants, as well as the consequences for later careers. Transition into the labor market, professional success, returns to higher education and gender differences are further topics central to the linkage project. Finally, the data allow to assess measures taken by the university to improve academic achievements by examining their impact on students' labor market success. The record linkage project thus provides various opportunities for research on the relationship between higher education and labor market outcomes.

As this paper focuses on the specific challenges and decisions in the context of the current linkage project, we do not provide a comprehensive overview of linkage techniques. For an in-depth coverage of the methods used here see, for instance, Christen (2012). We use a set of programs developed by the German Record Linkage Center (GRLC, see https://record-linkage.de). Antoni and Schnell (2019) provide a general overview of the GRLC's activities.

The remainder of this paper is structured as follows: Section 2 describes the two data sources and the non-unique linkage identifiers they contained. Section 3 provides information on the cleaning and standardising of these identifiers. Section 4 describes the comparison and classification steps of the linkage process. Section 5 discusses the linkage results, and compares characteristics of matched and unmatched graduates. Section 6 concludes with a short summary and some final remarks.

2 Data Sources

2.1 Student record data from the Nuremberg Institute of Technology (THN)

The graduate record data were provided by the academic controlling of the Nuremberg Institute of Technology (THN, https://www.th-nuernberg.de). THN is one of the largest Universities of Applied Sciences in Germany, with thirteen faculties and around 13,000 students. The 23,557 students in our data were born between 1952 and 2000. They were enrolled in one (or more) of the university's 74 different degree programs (19 diploma, 27 bachelor's and 28 master's programs) and graduated in the summer or winter semester between 2010 and 2020. During that time, 1,284 students earned a diploma degree, 18,487 students earned a bachelor's degree, and 6,373 students graduated from the THN with a master's degree, of which 3,844 already earned their bachelor's degree at the THN.

After the initial data preparation dealing with duplicate entries, it was possible to use the following linkage identifiers:

- First name, surname and, if available, birth name
- Day, month and year of birth
- Sex
- Address (zip code, place name, street name, house number)

Some students had more than one record within these data, which could be identified as belonging to the same person by an anonymous unique person identification number (*ID_linkage*) that was generated and assigned by the THN's academic controlling prior to the match. The data provided by the THN were transmitted to the IAB in two parts, which had to be appended before applying the pre-processing described in Section 3.

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* (DIM). These data originate 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 labor market policy measures

Data entries from any of these sources related to the same person are already integrated by the statistics department of the BA, where every individual represented in the data also receives a unique and time-consistent 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. Except for the birth name, which is not available in the records of the BA, the available linkage identifiers are identical to those of the THN.

To uphold the principle of data economy and to make the linkage process more efficient, we restricted the data extract from the records of the BA. The extract only contained linkage identifiers of people from the birth cohorts mentioned above and whose residential address had postcodes that actually occurred in the linkage identifiers of the THN.

3 Pre-processing

Although both data sources originate from data generating processes that already include some quality checks, errors and inconsistencies during the collection process cannot be completely ruled out. Therefore, the linkage identifiers described above may still include some errors. Using them for a comparison without any prior cleaning would thus lead to a suboptimal linkage result. We therefore used a set of pre-processing scripts developed by the GRLC to clean and standardize name and address fields. These routines are described in more detail by Antoni, Beckmannshagen, et al. (2023).

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

It was crucial to apply all steps of the pre-processing consistently on the linkage identifiers of both data sources to be linked to maximize comparability. Given that the THN provided their data in two separate files, we first appended them before consistently performing the rest of the pre-processing.

4 Comparison and classification

After the pre-processing of linkage identifiers we performed several steps to compare the record pairs from the two data sources. This section briefly describes the methods applied in these comparisons.

Deterministic linkage: In a deterministic linkage step, all or a predefined set of identifiers have to be fully identical for a record pair to be classified as a match. We perform this comparison as a set of simple *merge*-operations in Stata.

Blocking: Blocking limits the number of necessary comparisons of record pairs and thereby considerably reduces computing time. This is achieved by only comparing record pairs that have equal values on one or a set of blocking variables. As the deterministic linkage is computationally much less demanding than the distance-based linkage, we only apply blocking in the distance-based steps.

Distance-based linkage: In real-world data, a considerable share of records usually still has errors of some sort even after the pre-processing, e.g., typos, misspellings, inconsistent use of abbreviations, or different orderings of name components. In a deterministic linkage setting, even the smallest deviation between identifiers leads 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, we use Jaro-Winkler string comparator[1] for identifiers containing letters and an exact comparison for numerical identifiers. We use the Merge ToolBox (MTB) software[2] to perform the distance-based linkage.

Probabilistic linkage: The probabilistic linkage is an extension of the distance-based linkage in which the similarity of identifiers are not simply all added up. Identifiers are weighted according to their variance within the population under consideration. This is done because different identifiers vary in how strongly a given agreement is indicative for whether a record pair might actually be a true link. For instance, an agreement on the surname indicates the likelihood of a true match more strongly than an agreement on sex. Inversely, disagreement on the identifier 'sex' indicates a non-match more strongly than disagreement of the identifier 'surname' because of the latter's higher discriminatory power. See Table 1 for an overview of which comparative metric was used for which linkage identifier, and the m- and u-parameters applied in the probabilistic linkage.

Table 1: Overview of linkage identifiers, comparative metrics and parameters

Variable names	Comparative metrics	m-parameter	u-parameter
First name	Jaro-Winkler	0.801	0.002
Surname	Jaro-Winkler	0.85	0.0005
Birth date (day)	Exact (literally)	0.9	0.1
Birth date (month)	Exact (literally)	0.967	0.08
Birth date (year)	Exact (literally)	0.978	0.02
Sex	Exact (literally)	0.988	0.5
Zip code	Exact (literally)	0.889	0.049
Place name	Jaro-Winkler	0.876	0.012
Street name	Jaro-Winkler	0.792	0.001
House number	Exact (literally)	0.821	0.02

Array matching: When an identifier has more than one possible representation within one

[1] This string comparator determines the similarity of a pair of strings based on how many letters they share in the same position and in the same order. The Jaro-Winkler string comparator, a variant of the Jaro´s string comparator, puts a higher weight on letters at the beginning of strings.

[2] The MTB is maintained by the GRLC and can be downloaded and used for free for academic purposes. See https://record-linkage.de or Schnell, Bachteler, and Bender (2004) for more details on the software.

of the compared datsets, array matching can be applied to increase linkage success. In the project at hand, the THN data contain the surname and sometimes also the birth name for the same person, whereas the linkage identifiers from the BA only contain the surname at the time of data entry. An array match compares all representations of a given identifier for a record in the first dataset with all representations of that identifier in a record in the second dataset and only considers the highest similarity value as the result of the comparison. The function of array matching is not commonly available in record linkage software, but it is one of the options available in the MTB.

5 Linkage results

5.1 Linkage success rates

Table 2 summarizes the linkage steps we performed consecutively, starting with the strictest criterion of an exact agreement on all available identifiers (step *Deterministic 1*). This first step links persons based on the exact agreement of the first name, surname, sex, birth date (day, month, year), address (zip code, place name, street name, house number). In this step, 83.1 percent of the 23,557 persons available in the THN data extract are already linked successfully. The exact matching was then repeated iteratively by excluding one of the following (sets of) identifiers: zip code, place name, street name, house number, year of birth, month & day of birth. These six sub-steps are summarized in the step *Deterministic 2*. Relative to the first step, the cumulative contribution of these steps is rather small, as they only add 7.09 percentage points to the overall linkage rate.

Only persons not successfully linked in the deterministic steps are compared in the distance-based steps. To reduce computation time, blocking is used in all distance-based steps. *Distance-based 1*, which uses the 5-digit zip code as a blocking variable, adds 3.64 percentage points to the linkage rate. *Distance-based 2* uses blocking on month and year of the birth date as well as on sex, and it achieves additional 2.44 percentage points. Using the NYSIIS (New York State Intelligence and Identification System) phonetic code[3] on the first name in *Distance-based 3* added another 1.85 percentage points. The step *Distance-based 4* using blocking on the 3-digit zip code only added 0.07 percentage points to the linkage rate. Record pairs that could not be classified as matches throughout the prior steps were considered as potential matches. These cases were collected for a *Manual classification*, which in turn added another 0.06 percentage points to the overall linkage rate.

Of the 23,557 graduates from the THN data that could have been linked, we were able to find matches for 23,147 (98.26 percent). However, Table 3 shows that not all of these matches are available in the linked THN-IAB research data. For 196 of the matched cases, the Integrated Employment Biographies (IEB) of the IAB did not contain any records even though the Data Warehouse of the BA[4] did provide linkage identifiers for the underlying persons. This may

[3] NYSIIS is a phonetic code used to encode character strings such as names. See Taft (1970) for more details.

[4] We differentiate between the data from Data Warehouse of the BA, which contains the linkage identifiers originating from the sources listed in Section 2.2, and the research data in the IEB. The latter is created by IAB's department DIM based on extracts from the BA's Data Warehouse, but it only contains pseudonomized research data. See Schmucker, Seth, and vom Berge (2023) for more details on these research data.

Table 2: Summary of linkage steps

Linkage steps	N	share	description
	23557	100.00%	Number of persons in THN data before linkage
Deterministic 1	19577	83.10%	Exact agreement on first name, surname, sex, birth date (day, month, year), address (zip code, place name, street name, house number)
Deterministic 2	1671	7.09%	Exact agreement on identifiers as in *Deterministic 1* while excluding single identifiers in each of the six separate sub-steps (zip code; place name; street name; house number; year of birth; month & day of birth)
Distance-based 1	858	3.64%	Comparison of first name, surname, sex, birth date (day, month, year), address (zip code, place name, street name, house number); blocking on 5-digit zip code
Distance-based 2	575	2.44%	Comparison of first name, surname, birth date (day only), address (zip code, place name, street name, house number); blocking on year and month of birth and sex
Distance-based 3	436	1.85%	Comparison of first name, surname, sex, birth date (day, month, year), address (zip code, place name, street name, house number); blocking on NYSIIS-code of first name
Distance-based 4	16	0.07%	Comparison of first name, surname, sex, birth date (day, month, year), address (zip code, place name, street name, house number); blocking on 3-digit zip code
Manual classification	14	0.06%	Clerical review of all non-matches and manual classification of true matches.
Total	23147	98.26%	

Source: Date Warehouse of the BA, THN data; own calculations. *Note:* Steps *Deterministic 1* and *Deterministic 2* do not make use of the birth name available in the THN data. All distance-based steps use probabilistic linkage with the comparative metrics and parameters shown in Table 1.

Table 3: Linkage success and availability of research data

	N	share
Linkage candidates	23557	100.00%
Successfully linked	23147	98.26%
Successfully linked, IAB research data available	22951	97.43%
Successfully linked, IAB + THN research data available	22162	94.08%

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

be the case when a person has contact with a local employment agency, during which their address gets recorded (e.g., career counseling), but they have subsequently not (yet) entered any of the labor market states that are relevant for the IEB.

Additional 789 graduates that we could match during the linkage and for which IAB research data would have been available, are also not available in the linked THN-IAB research data. For those, the THN was able to provide linkage identifiers, but it was not possible to extract and prepare the corresponding THN research data for them. As a combined result of the linkage and the missing cases, the linked THN-IAB research data are available for 94.08 percent of the graduates that could have been linked.

5.2 Comparing Characteristics of Matched and Unmatched Students

Table 4 shows that there is a considerable similarity in most characteristics between the groups of graduates who were successfully matched and those who were not. Selected t-tests indicate that unmatched graduates' time to graduation was somewhat shorter, and their final grade point average (GPA) is slightly worse than the GPA of matched graduates. However, the difference between the two groups, especially regarding the final grade, has probably no economic relevance for most inquiries, despite its statistical significance. Furthermore, the difference in the average year of graduation suggests that earlier cohorts of graduates were harder to find in the administrative employment data. This might be due to the fact that earlier graduates are more likely to have changed their address or to have married and changed their surname between registering their information in the THN data and the time of the record linkage. Matched graduates are more likely to have participated in dual study programs than unmatched graduates. This is not surprising, given that dual study programs in Germany combine academic education with ongoing practical work experience. The latter usually entail an employment relationship with a company, which in turn leads to the student being registered in the social security data of the BA. We also find that the share of individuals with a foreign nationality is significantly higher among the unmatched graduates than among the matched graduates, suggesting that matching may have been impeded because some students in this group may have returned to their home countries after completing their studies instead of entering the labor market in Germany.

Chi-square independence tests shown in Table 5 also point to differences in some characteristics between matched and unmatched graduates. Concerning the type of university entrance qualification, unmatched graduates are more likely to have obtained their univer-

Table 4: Mean characteristics by match status, t-test of differences

	match	no match	diff.	t
No. of semester at graduation	10.377	9.753	-0.624	-3.724
University entrance qualification grade	2.581	2.536	-0.045	-1.305
Age at Graduation	26.414	26.260	-0.154	-0.783
Graduation grade	2.113	2.193	0.080	2.901
Graduation year	2015.944	2013.669	-2.274	-13.429
Dual students	0.093	0.051	-0.042	-2.743
Sex	0.436	0.439	0.003	0.129
Foreigner	0.079	0.233	0.154	10.728

Source: THN research data, own calculations based on 21,789 matched and 369 unmatched graduates. Due to a higher share of missing values, the results for the *University entrance qualification grade* are only based on 19,301 matched and 281 unmatched graduates. *Note:* The THN research data contain two different files, whose observations are pooled in Table 3. The analyses here only use data from one of these files, as it contains a higher number of observations and a more comprehensive set of variables.

sity entrance qualification abroad and they less frequently hold a general university entrance qualification compared to matched graduates. This aligns well with the fact that unmatched individuals are often foreigners compared to matched individuals, as shown in Table 4. Differences are also observed in the type of degree program. Unmatched students more frequently completed a diploma program than bachelor's and master's programs. Additionally, there are differences between matched and unmatched graduates in terms of the faculty in which they studied and the program they studied. The reason for that could be that international students (those with foreign nationality and foreign high school diplomas) mostly chose a few specific international (English-language) programs at the THN.

Table 5: Pearson χ^2-test

	results		
	chi2	df	p
Type of university entrance qualification	182.010	4	0.000
Faculty	21.788	12	0.040
Study program	297.866	51	0.000
Type of degree (BA, MA, Diplom)	117.251	2	0.000

Source: THN research data, own calculations based on 21,789 matched and 369 unmatched graduates. *Note:* The THN research data contain two different files, whose observations are pooled in Table 3. The analyses here only use data from one of these files, as it contains a higher number of observations and a more comprehensive set of variables.

6 Discussion

The overall linkage success rate was about 98 percent, which is way above the average of previous linkage applications with data of the BA.[5] Given that the administrative data of the BA only cover roughly 85 percent of the German workforce, the linkage can be classified as

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

very successful. Despite the high success rate, there are some differences between matched and non-matched graduates. Particularly, the sample of international graduates (those with foreign nationality or high school qualification from abroad) is somewhat underrepresented in the matched data. However, due to the high success rate of the matching, there are sufficient matched observations to allow analyses even for subgroups of graduates. Therefore, differences in characteristics are most likely negligible for most analyses.

Due to legal restrictions, the linked data cannot be made available to the general scientific community. The data can only be accessed for the purposes of replication and refereeing under the conditions described in the privacy policy Statements of the THN and IAB. Researchers interested in using the data are invited to contact us to explore a potential collaboration.

References

Antoni, Manfred, Nadine Bachbauer, Johanna Eberle, and Basha Vicari (2018). *NEPS-SC6-Erhebungsdaten verknüpft mit administrativen Daten des IAB (NEPS-SC6-ADIAB 7515)*. FDZ-Datenreport, 02/2018 (de).

Antoni, Manfred, Mattis Beckmannshagen, Markus M. Grabka, Sekou Keita, and Parvati Trübswetter (2023). *Survey data of SOEP-Core, IAB-SOEP Migration Sample, IAB-BAMF-SOEP Survey of Refugees and SOEP Innovation Sample linked to administrative data of the IAB (SOEP-CMI-ADIAB) 1975-2020*. FDZ-Datenreport 03/2023 (en). Nuremberg.

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

Antoni, Manfred and Rainer Schnell (2019). "The Past, Present and Future of the German Record Linkage Center (GRLC)." In: *Journal of Economics and Statistics* 239.2, pp. 319–331.

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

Comunian, Roberta, Sarah Jewell, and Alessandra Faggian (2017). "Graduate migration in the UK: an exploration of gender dynamics and employment patterns." In: *Graduate Migration and Regional Development*. Edward Elgar Publishing. Chap. 11, pp. 220–238.

Kaul, Ashok, Nathalie Neu, Anne Otto, and Manuel Schieler (2016). *Karrierestart, Mobilität und Löhne von Absolventen der Informatik*. IAB Regional 3/2016. IAB-Regional. Berichte und Analysen aus dem Regionalen Forschungsnetz. IAB Rheinland-Pfalz-Saarland.

Möller, Joachim (2017). *Regensburger Absolventenstudie. Rückblick auf das Studium und Einstieg ins Erwerbsleben von Regensburger Absolventinnen und Absolventen*. 1. Auflage. Literaturverzeichnis Seite [81]-83. Regensburg: Universitätsverlag.

Schmucker, Alexandra, Stefan Seth, and Philipp vom Berge (2023). *Sample of Integrated Labour Market Biographies (SIAB) 1975 - 2021*. FDZ-Datenreport 02/2023 (en). Nuremberg.

Schnell, Rainer, Tobias Bachteler, and Stefan Bender (2004). “A toolbox for record linkage.” In: *Austrian Journal of Statistics* 33, pp. 125–133.

Taft, Robert L (1970). *Name search techniques.* Bureau of Systems Development, New York State Identification and Intelligence System.

Teichert, Christian, Ingo Liefner, and Anne Otto (2021). “How wide is the gap? Comparing geography graduates’ labor market success with that of peers from business and computer science.” In: *Journal of Geography in Higher Education* 0.0, pp. 1–29.

Teichert, Christian, Annekatrin Niebuhr, Anne Otto, and Anja Rossen (2020). “Work experience and graduate migration: an event history analysis of German data.” In: *Regional Studies* 54.10, pp. 1413–1424.

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).**