Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Datenreport

Documentation of labour market data

03/2018
EN

Linked Inventor Biography Data 1980–2014

(INV-BIO ADIAB 8014)

Matthias Dorner,
Dietmar Harhoff,
Fabian Gaessler,
Karin Hoisl,
Felix Poege

Bundesagentur für Arbeit

Linked Inventor Biography Data 1980-2014 (INV-BIO ADIAB 8014)

Matthias Dorner
(Institute for Employment Research)
Dietmar Harhoff
(Max Planck Institute for Innovation and Competition, Ludwig-Maximilians-Universität München and CEPR)
Fabian Gaessler
(Max Planck Institute for Innovation and Competition and Technical University of Munich)
Karin Hoisl
(University of Mannheim, Copenhagen Business School, and Max Planck Institute for Innovation and Competition)
Felix Poege
(Max Planck Institute for Innovation and Competition, Munich Graduate School of Economics (MGSE) and IZA)

Dataset version:	INV-BIO ADIAB 8014 v1
	DOI: 10.5164/IAB.INV-BIO-ADIAB8014.de.en.v1
Documentation version:	INV-BIO-ADIAB8014_EN_v1_dok1
	DOI: 10.5164/IAB.FDZD.1803.en.v1

Die FDZ-Datenreporte beschreiben die Daten des FDZ im Detail. Diese Reihe hat somit eine doppelte Funktion: zum einen stellen Nutzerinnen und Nutzer fest, ob die angebotenen Daten für das Forschungsvorhaben geeignet sind, zum anderen dienen sie zur Vorbereitung der Auswertungen.

FDZ-Datenreporte (FDZ data reports) describe FDZ data in detail. As a result, this series of reports has a dual function: on the one hand, those using the reports can ascertain whether the data offered is suitable for their research task; on the other, the data can be used to prepare evaluations.

Abstract

This data report describes the Linked Inventor Biography Data 1980-2014 (INV-BIO ADIAB 8014), its generation using record linkage and machine learning methods as well as how to access the data via the FDZ.

Zusammenfassung

Dieser Datenreport beschreibt die verknüpften Erfinderbiografiedaten 1980-2014 (INV-BIO ADIAB 8014), deren Erstellung mittels Record Linkage und Machine Learning Methoden sowie den Datenzugang über das FDZ.

Keywords: inventors, labor market biography data, patent data, linked employer employee data, record linkage, data manual

Acknowledgements

The descriptions of the labor market biography and the establishment data draw on the SIAB data reports published by the IAB-FDZ (esp. the latest version by Antoni et al. 2016). Additionally, individual paragraphs were adopted and translated from IAB-internal data documentations by the Data and IT Management division (DIM) of the IAB. Further, we thank Stefan Seth, Dominik Braun, Kevin Ruf, Reinhard Sauckel and Oliver Senger for their practical assistance with the data throughout the project. Jesper Zedlitz provided helpful historical location data from the Geschichtliches Ortverzeichnis (GOV). Philipp vom Berge (IAB) and Manfred Antoni (IAB) provided helpful comments for setting up the research data set.

Funding

The production of the INV-BIO ADIAD 8014 research data was supported by the DFG (Deutsche Forschungsgemeinschaft) in the project "Produktivitätseffekte von Erfindermobilität in Agglomerationen und in Teams" (Grant Number 329144242).

Data availability and access

The dataset described in this document is available only for non-commercial research in academic organizations. Further information on the data access are provided on the website of the FDZ at http://fdz.iab.de.

Table of content

1 Introduction and outline

1.1 Introduction

This data report describes the linked inventor biography data set INV-BIO ADIAB 1980-2014 (henceforth also: INV-BIO), which is based on inventor and patent information obtained from patent register data that are linked to administrative labor market career data on individuals and their employing establishments.

The data set comprises 152,350 inventors who are listed on patent filings at the European Patent Office (EPO) between 1999 and 2011 and who are successfully linked with employees in that time period.

Linking inventors with individual level data that contain a unique person identifier (ID), such as employee records in the social security register based IAB data, represents a disambiguation approach alternative to those currently performed in the literature towards generating consistent inventor biography track records.

The INV-BIO data set records complete biographies of unique inventors from 1980 until 2014. For this period, inventor patent track records based on patent registers of the EPO and the German Patent and Trademark Office (DPMA), and labor market biographies originating from social security data are combined in a research data set. High quality linked data like these have so far not been available for the study of inventor careers or invention processes (more generally: knowledge production) in the context of the labor market. In addition, the national innovation system of Germany provides an interesting case for studying inventive processes, because Germany as the country of origin of patents, is ranked amongst the top group of countries with respect to the protection of intellectual property by patents.

Against the background of the rise of the knowledge economy and the increasing importance of innovation for economic growth, economic research in recent years has increasingly focused on the role of individuals, their incentives and behavior for the generation innovative output, as well as the contexts in which innovations are being generated and how to manage these environments. Although innovation is a much broader concept than what is defined in legal terms as a patent, inventions described in patents share important characteristics[1] of what scholars in the field of science and innovation consider innovation, making patents a frequently used proxy for the intangible variable (see Griliches 1990). Hence, the biographies of inventors comprised in these data will be a highly informative source for research and may contribute to a better understanding of these research objectives. In recent years, scholars in several

[1] Patents, in order to be granted, must describe the solution a technical problem, which is novel and involves an inventive step, i.e. the invention must be non-obvious and go beyond the current state of the art. Moreover, the invention must be commercially applicable.

countries have made substantial efforts towards linking inventors with census data (Finland, Sweden, US), social security data (Italy) or tax records (US) to study the antecedents and social returns to invention.[2]
Data generated in these projects, however, are not freely available to other researchers. This inventor biography research data set thus is the first to be used for scientific non-commercial researchers. Furthermore, it is the first data set to describe the generation of knowledge and intellectual property by individuals in the German labor market and national innovation system.

The INV-BIO data set was generated at the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the IAB in cooperation with the Max Planck Institute for Innovation and Competition and funded by the Deutsche Forschungsgemeinschaft (DFG) (Grant Number 329144242).

1.2 Sampling frame

The sampling frame of the INV-BIO data is the population of inventors who are listed in the PATSTAT[3] data on patent applications filed with the European Patent Office (EPO) between 1999 and 2011 and reside in Germany. These data alone, however, do not provide a unique identifier of individuals across the 699,000 patent applications.[4]
The 152,350 unique inventors recorded in the INV-BIO data were identified from the population using a unique methodological approach combining record linkage and techniques of machine learning. This approach enabled to identify inventors in the social security data from 1999 to 2011 in the first step, and in the second step, to assign patent records to the previously identified unique inventor accounts based on predictive methods from the machine learning toolkit. Consequently, unique identifiers in the form of pseudonymized social security IDs are only available for the subpopulation of linked inventors but not the rest of the population who were not identified in the labor market data. This group of unmatched inventors are either not found because of data quality issues or they belong to one group of individuals that are not covered by the IAB employment data, such as self-employed persons, freelancers, civil

[2] See Toivanen and Väänänen (2012), Depalo and Di Addario (2014), Jung and Ejermo (2014), Aghion et al. (2017), Akcigit et al. (2017) and Bell et al. (2017).
[3] PATSTAT contains bibliographical and legal status patent data from leading industrialized and developing countries. This is extracted from the EPO's databases
[4] The lack of an individual identifier across patents complicates the analysis of inventor data. For instance, inventors with the same common name listed on two patents may not be the same person. Further, organizational and/or geographical mobility as well as typographical inconsistencies in the filing of patents (even by the same inventor/applicant organization) complicate the identification of unique persons across patents.

servants or students. According to the PatVal Survey of European inventors, these groups should account for about 10 percent of the population (Gambardella et al. 2005).

Evaluated at the level of patent families, the INV-BIO data include 643,856 patent families (DOCDB definition of the EPO) and represents approximately 71.4 percent of the inventions in Germany during the time window from 1999 until 2011 (a more detailed evaluation and discussion is provided in the Appendix).
For comparison at the person level, the average number of individuals that are reported by their employing establishments as employees in research and development in the period 1999-2011 amounts to 637,308 individuals. Evaluated at this number, the linked inventors contained in the INV-BIO data set correspond to roughly 24 percent of this potentially patenting population of employees. A more thorough evaluation of the representativeness at the person level, however, is not possible because of the inventor disambiguation problem in the patent register.

1.3 Data use

1.3.1 Data access and application

The INV-BIO data are weakly anonymous data with pseudonymized identifiers for individuals, employers and patents, but otherwise represent original data as recorded in the underlying register data sources. Therefore, it is only allowed to process and analyze these data in the context of a research visit at one of the official Research Data Centre (FDZ) sites[5] and by subsequent remote data execution using the FDZ infrastructure.

To access the INV-BIO research data, it is first necessary to submit an application to the FDZ. This application should include a description of the planned research project comprising a statement on the relevance with respect to labor market and occupational research as well as details on the required variables or data modules. FDZ staff reviews the application. Research projects are only accepted when the purpose of research is labor market and/or occupational research, as defined by the Social Code §272 SGB X. Other research projects will be rejected. Moreover, the FDZ is strictly bound to the principle of data economy. Therefore, applications for the INV-BIO data must include arguments why the INV-BIO data set is actually required for the outlined research project and that not any other scientific research data set can be used. To this end, in their applications researchers should discuss patent register data such as PATSTAT or the OECD REGPAT data base as alternative sources in the context of inventor

[5] For an up-to date list of sites, please check with the FDZ website at http://fdz.iab.de.

information. At the firm level, data sets including patent information such as Bureau Van Dijk ORBIS or (aggregate) survey data such as the Community Innovation Survey are references that should be checked and/or discussed as potential alternative data.
When approval for the project has been eventually granted by the FDZ, a data use agreement is concluded (via the FDZ) between the IAB and the researcher's institution. Further details on the application process and on data processing are provided on the FDZ website at https://fdz.iab.de/en/FDZ_Data_Access.aspx.

1.3.2 Data structure and management

The INV-BIO data are documented bilingually and include both German and English labels[6]. The data have a modular structure and are organized in three separate files to facilitate data processing and to save memory in the data management system. The three files, the Inventor patent file (INV-PAT), the Labor market biography file (INV-SIAB) and the Basis Establishment file (INV-BHP) are briefly described below.

First, the patent register based information on inventors and patents are recorded in the **Inventor patent file (INV-PAT)** that is provided by the Max Planck Institute for Innovation and Competition. These data are structured by the inventor (individual ID) and the patent ID (which is at the patent family level).[7] As such, the file records inventors that were successfully linked to an individual in the IAB data and all patented new technical content on inventions to which the unique inventors contributed between the years 1980 and 2014. Moreover, the inventions described therein are characterized in detail using technology and bibliographical information as recorded in patent registers. Note that patent information is duplicated in this data structure as patents may be linked to more than one single inventor.

Second, the linked-employer-employee data describing the labor market careers of inventors are provided by the IAB. These data are organized in a **Labor market biography file (INV-SIAB)** and a complementary **Basis establishment file (INV-BHP)**. Both files mirror the structure of the Sample of the Integrated Labor Market Biographies (SIAB), which are provided by the IAB-FDZ as a 2 percent random sample of the population of individuals in the underlying social security register data, plus complementary establishment records. The INV-SIAB differs

[6] The Stata commands label language en or label language de can be used to change the language of the labels between English (en) or German (de), respectively.
[7] A simple patent family, also known as DOCDB family, is a collection of patent documents that are considered to be covering one single invention in one or several jurisdictions. The technical content covered by the applications in one simple patent family is considered to be identical. Patent applications that are members of a given DOCDB patent family will have exactly the same set of priority filings. For more details on the definition of patent families at the EPO (EPO 2017). A more comprehensive discussion of the concept of patent families is provided in Martinez (2010).

from the SIAB sample in the scope of variables included in the data. While the set of employment variables is also recorded in the INV-SIAB, information on benefits, job search and measures is reduced in scope and level of detail to avoid highly singular data.

The **Labor market biography file (INV-SIAB)** is structured by the individual person identifier (ID) and episodes of employment, registered unemployment or job search. Hence, it is similar to the SIAB data set. The set of variables includes socio-demographic characteristics of the individuals, information on employment, benefit receipt and job search activity, variables describing the residential location and technical variables. Establishment IDs for employment records are provided to combine the data with additional information on the employing establishment of the focal inventor.

The **Basis establishment file (INV-BHP)** is structured by establishment ID and year. It comprises variables recording the location of the site, the industrial classification of economic activities in the NACE scheme as well as basic structural information on the establishment as of the reference date of 30 June in each year.
The INV-BHP includes only a subset of the rich establishment level data available at the FDZ. On request, the basic information provided for establishments may be complemented with other data from the Establishment History Panel (BHP) data of the FDZ, that include more comprehensive structural information on establishments, establishment dynamics variables (year and type of establishment entry/exit) as well as annual worker flow accounts for the sites.

The names of the files that are included in the INV-BIO data and provided by the FDZ are listed below:

Inventor patent file (INV-PAT)

- INV-BIO_ADIAB_8014_v1_inv-pat_7514_v1.dta

Labor market biography file (INV-SIAB)

- INV-BIO_ADIAB_8014_v1_siab.dta

Basis Establishment File (INV-BHP)

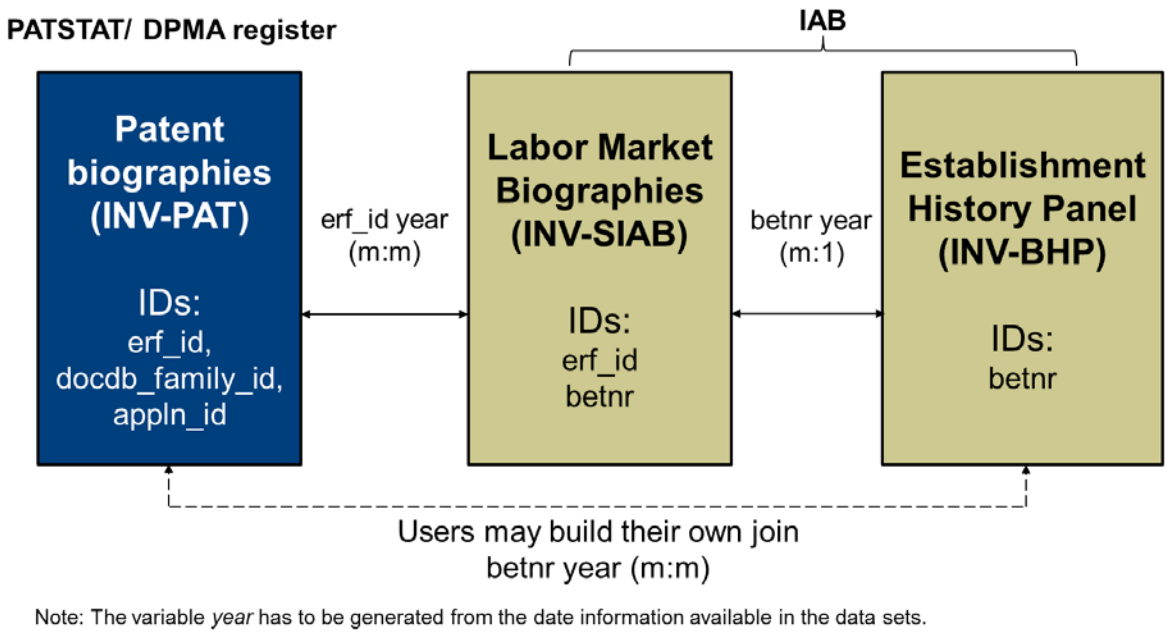
- INV-BIO_ADIAB_8014_v1_bhp_basis_v1.dta
    - **BHP extension files available on request**
        i. Detailed establishment characteristics (structured by variable blocks)
        INV-BIO_ADIAB_8014_v1_bhp_v1_YYYY.dta, with YYYY = 1980 - 2014
        ii. Worker flows
        INV-BIO_ADIAB_8014_v1_bhp_inflow_v1.dta
        INV-BIO_ADIAB_8014_v1_bhp_outflow_v1.dta

iii. Establishment dynamics

INV-BIO_ADIAB_8014_v1_bhp_entry_v1.dta

INV-BIO_ADIAB_8014_v1_bhp_exit_v1.dta

Linking information across the files will be required for most scientific analyses of the INV-BIO. Figure 1 depicts the data base structure and documents the merging keys required for combining the files organized in the INV-BIO data.



Note: The variable *year* has to be generated from the date information available in the data sets.

Figure 1: Data base structure of the INV-BIO

In the following, the merging keys are briefly described. Detailed example code for the two merges to be used with Stata is provided in the Appendix. First, the INV-PAT file can be merged with the INV-SIAB to relate patent characteristics to the employment records in the linked employer-employee data of the IAB. This merge relies on the individual ID in combination with the time information available in the data. To perform this merge between data structured in (parallel) employment and non-employment episodes (spell data) in the labor market biography data and the event information (quarterly data) in the inventor patent data, the structure of the data sets has to be considered, i.e. structural transformations of the data have to be performed (for Stata code examples, see Appendix A2).

The INV-SIAB and the establishment records of the INV-BHP can be merged on the establishment ID (betnr) and year (of the employment episode). The year variable is not readily available in the biography data file but has to be generated by the researcher in the two files

prior to the merge. This merge enables researchers to enhance the information on employment spells with a rich set of structural characteristics of the employing establishment, NACE industry classification or location information of the site. A visualization of the merge between the labor market biography data and the establishment data available at the FDZ is presented in Figure 2.

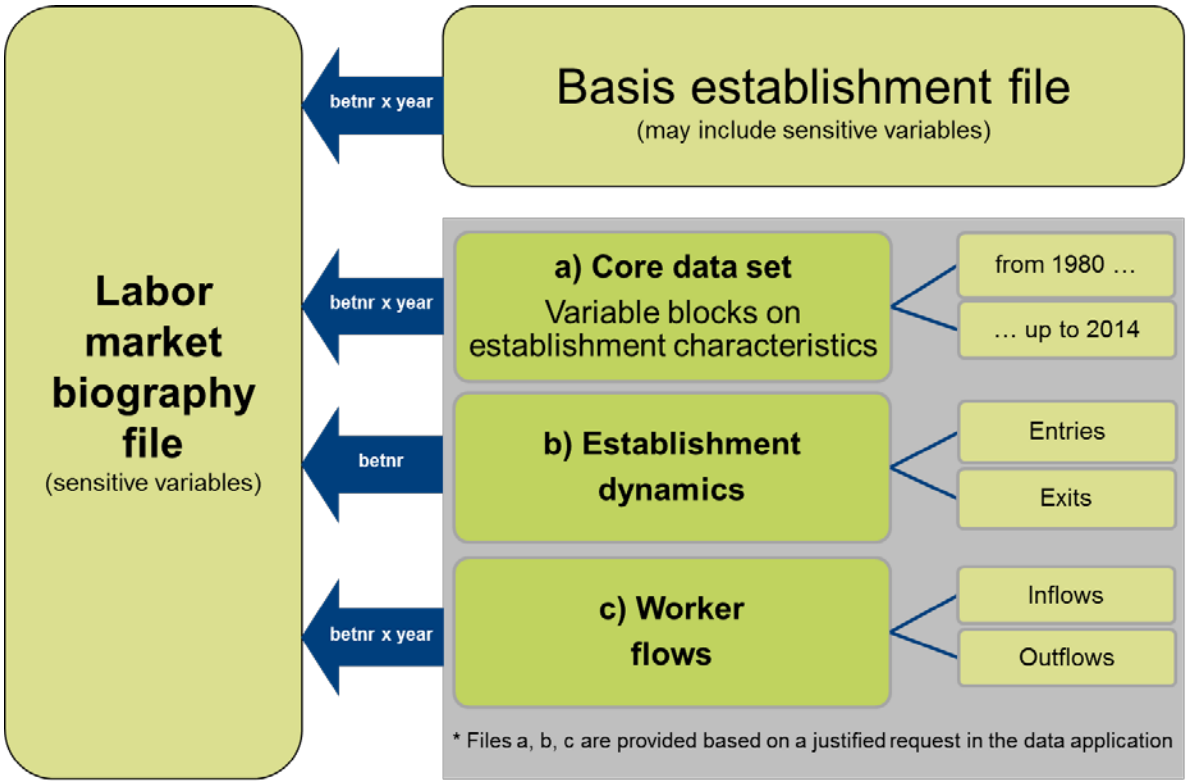


Figure 2: Data storage of individual and establishment file data

1.3.3 Data privacy and anonymization

The INV-BIO data are pseudonymized data, i.e. explicit identifiers from the original data sources such as names or address information about the individuals and their employing establishments are replaced with pseudonymized identifiers that do not allow the identification of inventors in public patent data.

Unique individuals may be tracked over the course of their career in the data based on the individual ID, which is aligned to the social security ID. Similarly, the establishment ID is a pseudonymized version of the original establishment identifier that is recorded as a unique entity over time in the establishment register of the Federal Employment Agency (BA).

To protect the privacy of the individuals and establishments in the data and to limit the threat of potential deanonymization, some information from the original register are classified as

sensitive and/or subject to further data anonymization. For instance, the information on work and residential location of the employee are actually available in the register at very detailed address level that easily deanoymize the underlying individual or firm. Using the hierarchical levels of the official territorial system classification in Germany (Amtlicher Gemeindeschlüssel, AGS) and the NUTS scheme of Eurostat, we provide only coarsened location information that enables locating both individuals and establishments in the territorial structure of cities and districts (NUTS 3 level), but not in the detailed geographical structure of municipalities (within districts). The same logic applies to occupational or industrial classifications, which are anonymized or classified accordingly. Information on nationality or the dates from which establishment entry/exit can be inferred are available in two versions, first, as is with full detail and, second, in a coarsened form. A full list of sensitive variables that are subject to anonymization is provided below.

Labor market biography file (INV-SIAB):

- Nationality (nation)
- Occupational classification:
    - KldB 1988, occupation group – current/ most recent, 3-dig. level (beruf1988_3)
    - KldB 2010, occupation group current/ most recent, 3-dig. level (beruf2010_3)
- Place of residence: district (Kreis/ NUTS 3) (wo_kreis)

Basis Establishment file (INV-BHP):

- Place of work: district (Kreis) (ao_kreis)
- Industry classification:
    - WS 1973 / German classification of economic activity 1973 (3-digit level) (w73_3)
    - NACE Rev. 1/ German classification of economic activity 1993 (3-digit level) (w93_3)
    - NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993 generated – completed by extrapolation / imputation (w93_3_gen)
    - NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993 generated – type of completion (group_w93_3)
    - NACE Rev. 1.1/ German classification of economic activity 2003 (3-digit level) (w03_3)
    - NACE Rev. 2/ German classification of economic activity 2008 (3-digit level) (w08_3)

- NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008 generated – completed by extrapolation / imputation (w08_3_gen)
- NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008 generated – type of completion (group_w08_3)

Inventor patent file (INV-PAT)

- Patent technology fields TF34 (area34)

Variables classified as sensitive by default are either omitted or only included in the data in a coarsened version. Access to the more detailed content of the classified variables will be provided only if the variable has been approved in the official data application process prior to the start of the research. For further information and advice on the application process for sensitive variables, please contact the FDZ.

1.4 Changes as compared to previous versions

Does not apply.

1.5 Outline

CONTENT CHARACTERISTICS	
Topics/ groups of variables	a) Inventor patent file (INV-PAT) Patent track records of unique inventors including detailed patent characteristics (DOCDB family level) describing the invention recorded in the data. b) Labor market biography file (INV-SIAB) Employee History (Beschäftigten-Historik - BeH): Annual notifications and end-of-employment notifications submitted to the social security agencies for employees covered by social security and employees in marginal part-time employment. Benefit Recipient History (Leistungsempfänger-Historik - LeH): Information on benefit receipt in accordance with Social Code Book III (SGB III) for recipients of unemployment benefit, unemployment assistance and maintenance allowance. Unemployment Benefit II Recipient History drawn from A2LL (Leistungs-Historik-Grundsicherung aus A2LL - LHG): Data on individuals in receipt of basic social security benefits in accordance with Social Code Book II (SGB II) (Types of institution: cooperation of employment agencies and municipalities/joint facilities, separated responsibilities/municipalities exercising their duties separately, authorized municipalities) Jobseeker History (Arbeitsuchenden-Historik - ASU):

	Information on job search activity Jobseeker History drawn from XSozial-BA-SGB II (XASU): Information on job search activity reported via the transmission standard XSozial-BA-SGBII to the BA by authorized municipalities. Participants-in-Measures History File (MTH): Information on participation in employment and training measures (not including measures of the authorized municipalities)
Data unit	Register based biographical records on inventors who are listed on at least one patent with the European Patent Office 1999-2011 and who were linked to a unique employee recorded in administrative labor market data of the IAB in the same period. For all inventors, the full biographies (1980-2014) are recorded in the data.
Number of cases	a) Inventor patent data (INV-PAT): 1,209,687 patent-inventor records - 152,350 unique inventors/employees - 643,856 patents (DOCDB family level) b) Administrative labor market data (INV-SIAB, INV-BHP): 152,350 unique inventors - 4,592,998 original observations - 4,957,835 non-overlapping observations (after episode splitting) 148,965 establishment IDs - thereof: 144,646 (97.1%) in INV-BHP establishment file
Period covered	The period covered depends on the data source. a) Inventor patent file (INV-PAT) (appln_filing_date) 1 January 1980 - 31 December 2014 b) Labor market biography file (INV-SIAB) BeH: 1 January 1980 - 31 December 2014 LeH: 1 January 1980 - 31 December 2014 ASU: 1 January 1997 - 31 December 2014 LHG: 1 January 2005 - 31 December 2014 XASU: 1 January 2005 - 31 December 2014 MTH: 1 January 2000 - 31 December 2014
Time reference	a) Inventor patent file (INV-PAT) Quarterly information on DOCDB patent families b) Labor market biography file (INV-SIAB) Episodes exact to the day c) (Basis) Establishment file (INV-BHP) Establishment characteristics at annual reference date June 30
Territorial system	Amtlicher Gemeindeschlüssel (AGS)/ NUTS classification of Eurostat: federal states (Bundesland/ NUTS 1), districts (Kreise und kreisfreie Städte/ NUTS 3).
Date of territorial system	401 districts (Kreise and Kreisfreie Städte/ NUTS 3) and 16 federal states (Bundesland/ NUTS 1) as of 31 December 2016

METHODOLOGICAL CHARACTERISTICS	
Data design	Identification of inventors in the integrated employment biographies of the IAB using methods of record linkage.
Data sources	a) Inventor patent file (INV-PAT) PATSTAT data base (10/2017) and register data of the German Patent- and Trademark Office; Data and variables were processed at the Max Planck Institute for Innovation and Competition and at the IAB. b) Labor market biography file (INV-SIAB) and (basis) establishment file (INV-BHP) Social insurance agencies, Federal Employment Agency; Both data sources were further processed at the IAB.
Update frequency	No updates of the data set planned
File format and size	Data sets of INV-BIO ADIAB 8014 are provided in Stata (*.dta) format only a) Inventor patent file (INV-PAT): 80 MB b) Labor market biography file (INV-SIAB): 241 MB c) (Basis) Establishment file (INV-BHP): 26 MB Note: file size information excluding sensitive variables.
File organization	The INV-BIO ADIAB 8014 data are stored in three separate files. Additional information at the level of establishments may be requested following a justified application.
DATA ACCESS	
Data access	On-site use at the FDZ of the BA at the IAB and subsequent remote data execution.
Degree of anonymization	Weakly anonymous
Sensitive variables:	a) Inventor patent file (INV-PAT) Technological field of patent (area34) b) Labor market biography file (INV-SIAB) Nationality (nation), occupation group (beruf1988_3, beruf2010_3), residential location: districts (wo_kreis) c) (Basis) Establishment file (INV-BHP) Work location: districts (ao_kreis),

	Industry classifications: WS 1973/ German classification of economic activity 1973 (3-digit level) (w73_3), NACE Rev. 1/ German classification of economic activity 1993 (3-digit level) (w93_3), NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993 generated – completed by extrapolation / imputation (w93_3_gen), NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993 generated – type of completion (group_w93_3), NACE Rev. 1.1/ German classification of economic activity 2003 (3-digit level) (w03_3), NACE Rev. 2 / German classification of economic activity 2008 (3-digit level) (w08_3), NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008 generated – completed by extrapolation / imputation (w08_3_gen), NACE Rev. 2/ WZ 2008German Classification of Economic activity 2008 generated – type of completion (group_w08_3)
Citation of data and data documentation	Data: "The data basis of this paper is the linked inventor biography data 1980-2014 (INV-BIO ADIAB 8014). These data were accessed on-site at the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) and/or via remote data access at the FDZ." Citation of data documentation: Dorner, M.; Harhoff, D.; Gaessler, F.; Hoisl, K.; Poege, F. (2018): Linked Inventor Biography Data (INV-BIO ADIAB) 1980-2014. FDZ-Datenreport, 03/2018 (en), Nuremberg. DOI:10.5164/IAB.FDZD.1803.en.v1
Dataset version	Linked Inventor Biography Data (INV-BIO ADIAB) 1980-2014 v1. DOI: IAB.INV-BIO-ADIADB8014.de.en.v1

Table 1: Data outline

1.6 List of variables

The overview provided in Table 3 lists complete variables names and the corresponding descriptions. It also tabulates the availability and completeness of the data by source:

	Variable is available for the data source. Degree of completeness always > 0.85.
	Variable is available for the data source. Lower or varying degree of completeness, see description of variable and frequency count.
	Variable is not available for this data source. Degree of completeness always < 0.05.

Table 2: Degrees of completeness of the variables

Examples: First, the variable ‘daily wage, daily benefit rate’ (tentgelt) is only available for BeH and LeH observations; the observations of the other data sources contain the missing value “.n” for this variable. Second, some variables have different contents depending on the data source. For instance information on the ‘employment status’ (erwstat) originating in the BeH source contain the status as recorded in the official employment notification procedure. For records originating in the LeH, the type of benefit is recorded. LHG and XLHG contain the SGB II status, ASU and XASU records the job search status and, finally, for MTH observations, the measure type is documented in the variable. These differences are not obvious from the label of every variable but are described in detail in the variable descriptions.

List of variables	Page	INV-SIAB/ INV-BHP						INV-PAT
		BeH	LeH	LHG	ASU	XASU	MTH	
Identifiers								
Inventor ID (erf_id)	45	■	■	■	■	■	■	■
Establishment ID (betnr)	45	■						
Generated technical variables								
Observation counter per person (spell)	47	■	■	■	■	■	■	
Source of spell (quelle)	47	■	■	■	■	■	■	
Year (jahr)	47	■	■	■	■	■	■	
Period of validity								
Original start date of observation (begorig)	47	■	■	■	■	■	■	
Original end date of observation (endorig)	48	■	■	■	■	■	■	
Start date of split episode (begepi)	48	■	■	■	■	■	■	
End date of split episode (endepi)	49	■	■	■	■	■	■	
Personal information								
Gender (frau)	49	■	■	■	■	■	■	
Year of birth (gebjahr)	49	■	■	■	■	■	■	
Nationality (nation) *	49	■	■	■	■	■	■	
Nationality, aggregated (nation_gr)	50	■	■	■	■	■	■	
Marital status (famst)	50		■					
Number of children (kind)	50		■					
Vocational training (ausbildung)	51	■			■	■	■	
School leaving qualification (schule)	53	■			■	■	■	
Information on employment, benefit receipt and job search								
Daily wage, daily benefit rate (tentgelt)	54	■	■					

List of variables	Page	INV-SIAB/ INV-BHP						INV-PAT
		BeH	LeH	LHG	ASU	XASU	MTH	
KldB 1988, Occupation main group – current/most recent (beruf1988_2)	55							
KldB 1988, Occupation group – current/most recent (beruf1988_3) *	56							
KldB 2010, Occupation main group – current/most recent (beruf2010_2)	57							
KldB 2010, Occupation group – current/most recent (beruf2010_3) *	58							
KldB 2010, Level of requirement – current/ most recent (niveau)	59							
Part-time (teilzeit)	59							
Employment status (erwstat)	60							
Transition zone (gleitz)	62							
Temporary agency work (leih)	62							
Fixed-term contract (befrist)	62							
Reason of cancellation/ notification/ termination (grund)	63							
Start date of unemployment (alo_beg)	64							
Duration of unemployment (alo_dau)	64							
Location data								
Place of residence: district (Kreis/ NUTS 3) (wo_kreis)*	65							
Place of residence: federal state (Bundesland/ NUTS 1) (wo_bula)	66							
Establishment characteristics								
WS 1973 German Classification of Economic activity 73, 2-digit level (w73_2)	66							
WS 1973 German Classification of Economic activity 73, 3-digit level (w73_3) *	67							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, 2-digit level (w93_2)	68							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, 3-digit level (w93_3) *	68							
NACE Rev. 1.1 / WZ 2003 German Classification of Economic activity 2003, 2-digit level (w03_2)	69							
NACE Rev. 1.1 / WZ 2003 German Classification of Economic activity 2003, 3-digit level (w03_3) *	70							

List of variables	Page	INV-SIAB/ INV-BHP						INV-PAT
		BeH	LeH	LHG	ASU	XASU	MTH	
NACE Rev. 2 /WZ 2008 German classification of classification Economic activity 2008, 2-digit level (w08_2)	70							
NACE Rev. 2 /WZ 2008 German classification of Economic activity 2008, 3-digit level (w08_3) *	71							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, time consistent, 2-digit level (w93_2_gen)	72							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, type of imputation (group_w93_2)	72							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, time consistent, 3-digit level (w93_3_gen) *	73							
NACE Rev. 1/ WZ 1993 German Classification of Economic activity 1993, type of imputation (group_w93_3) *	73							
NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008, time consistent, 2-digit level (w08_2_gen)	74							
NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008, type of imputation (group_w08_2)	74							
NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008, time consistent, 3-digit level (w08_3_gen) *	75							
NACE Rev. 2/ WZ 2008 German Classification of Economic activity 2008, type of imputation (group_w08_3) *	75							
Year of first appearance of establishment (grd_jahr)	76							
Year of last appearance of establishment (lzt_jahr)	76							
Total number of employees (az_ges)	77							
Number of full-time employees (regular workers + others) (az_vz)	77							

List of variables	Page	INV-SIAB/ INV-BHP						INV-PAT
		BeH	LeH	LHG	ASU	XASU	MTH	
Number of employees in marginal part-time employment (az_gf)	77							
Mean imputed wage all full-time employees (te_imp_mw)	78							
Place of work: district (Kreis/ NUTS 3) (ao_kreis) *	78							
Place of work: federal state (Bundesland/ NUTS 1) (ao_bula)	79							
Patent characteristics								
DOCDB family ID (docdb_family_id)	79							
Application ID, earliest matched appl. within DOCDB family by inventor (appln_id)	80							
Number of applications within DOCDB family (docdb_family_size)	80							
Earliest application filing date within DOCDB family (earliest_filing_date)	80							
Application filling date, earliest matched appl. within DOCDB family by inventor (appln_filing_date)	81							
Earliest publication date within DOCDB family (earliest_publn_date)	81							
Patent application is granted (granted)	81							
Grant date of patent application (grant_date)	82							
DOCDB family citations within {X} yrs – DE (cit_docdb_DE_Xyrs)	82							
DOCDB family citations within {X} yrs – EP (cit_docdb_EP_Xyrs)	82							
DOCDB family citations within {X} yrs – US (cit_docdb_US_Xyrs)	83							
Generality measure based on forward citations to DOCDB (generality_docdb)	83							
Originality measure based on backward citations of DOCDB (originality_docdb)	83							
Number of inventors (nb_inventors)	84							
Foreign inventors (d_foreign_inv)	84							
Complete inventor team (pat_complete)	85							
Number of applicants (nb_applicants)	85							

List of variables	Page	INV-SIAB/ INV-BHP						INV-PAT
		BeH	LeH	LHG	ASU	XASU	MTH	
Foreign applicants (d_foreign_appl)	85							
Technology area (area34)	86							
Technology main (mainarea34) *	86							
Matched inventors are employed with multiple establishments (multi_betnr)	87							
Average distance between matched inventors (mean_dist_inv)	87							

Note: Sensitive variables indicated by asterisk (*)

Table 3: List of variables with degree of completeness

1.7 Volume structure

No. of records		before splitting	after splitting
Labor market biography file (INV-SIAB)			
BeH		4,282,272	4,500,194
LeH		130,495	169,602
LHG		5,869	15,612
ASU		150,190	235,280
XASU		1,531	2,678
MTH		22,641	34,469
Total number of episodes		4,592,998	4,957,835
Inventors (unique)	152,350		
Establishments (unique) - thereof in INV-BHP	148,965 144,646		
DOCDB patent families (unique)	643,856		

Table 4: Volume structure

2. Data sources

2.1 Patent register data

The patent information included in the INV-BIO data and used for the generation of the data are drawn from register data recorded in PATSTAT, which contains bibliographical, procedural and legal status data on patent applications handled by the European Patent Office (EPO), and from DPMAregister, the online patent register of the German Patent and Trademark Office (DPMA).

The PATSTAT data represent a collection of tables that document the characteristics and events of patents along their life cycle and that are of legal importance to administer the

invention in the intellectual property rights system. The European Patent Register database is the core of the PATSTAT universe. It is released twice a year as a fully integrated database and contains bibliographic, legal and procedural information on published European patent applications and on published Patent Cooperation Treaty (PCT) applications for which the EPO is a designated office (so called Euro-PCT applications). This database is extracted from the European Patent Register, which stores all publicly available information the EPO has on European patent applications as they pass through the application and examination procedure. The information includes among others: the names of applicants, inventors, opponents and legal representatives, procedural events during application and examination proceedings, opposition and appeal proceedings.

The EPO Worldwide Legal Status database, as the second major source, is the most comprehensive source of information on legal events occurring during the life of a patent, before or after grant (e.g. change of ownership, withdrawal of the application, entrance into the national phase). The records in this table originate from the patent gazettes and registers of various national patent authorities, including the EPO and World Intellectual Property Organization (WIPO), which administers international patent applications filed under the Patent Cooperation Treaty (PCT). Currently over 50 offices, including the DPMA, provide the EPO with legal status data.

External complementary data files generated by the OECD, the KU Leuven and the WIPO such as name harmonization, additional indicators for patents or technology concordance tables can be used to enhance the raw data in PATSTAT.[8]

We enhance the PATSTAT data extract with information drawn from the DPMAregister that is exclusively recorded therein and refers to national patent applications (DE) that are neither passed on to the EPO nor filed under the PCT route.

The Max Planck Institute for Innovation and Competition in Munich licensed the patent register data used for the generation of the dataset. The required data tables on inventors, applicants and addresses were made available to link inventors with employees in the IAB register data. Variables describing the patents such as citation counts or technology classifications were generated at the Max Planck Institute for Innovation and Competition and provided for the research data set.

[8] For an overview of complementary databases to PATSTAT, see http://documents.epo.org/projects/babylon/eponet.nsf/0/AD3685345749FB6EC12581FA00589502/$File/patstat_complementary_databases_v1.0_en.pdf, accessed 2018-07-20.

2.2 Administrative data from the IEB

The administrative individual data were drawn from the Integrated Employment Biographies (IEB) of the IAB. They unite data from five different data sources, each of which may contain information from different administrative procedures. In addition, some supplementary variables from these data sources, which are not part of the IEB, are incorporated into the administrative individual data. Figure 3 illustrates the data streams leading to the labor market biography file (INV-SIAB). The Basis Establishment file (INV-BHP) is the corresponding subset of the full BHP data to enhance the biographies of the inventors recorded in the INV-BIO research data.

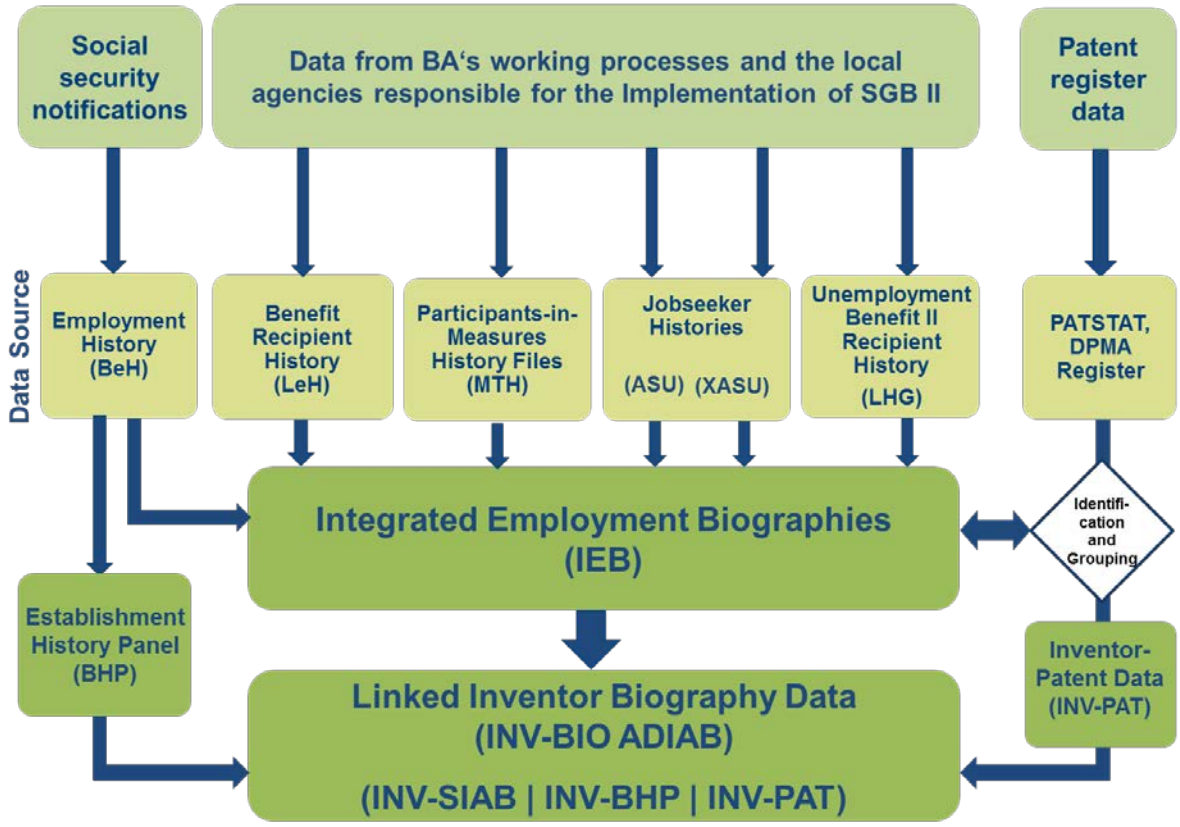


Figure 3: Data sources of the INV-BIO ADIAB 8014

2.2.1 Employee History (BeH)

The source of data regarding employment is the Employee History (Beschäftigten-Historik - BeH) of the IAB. The data basis is the integrated notification procedure for health, pension and unemployment insurance, which came into effect as of 1 January 1973 (and was extended to cover eastern Germany as of 1 January 1991) and is known by the abbreviation DEÜV (previously DEVO / DÜVO) (for further details see: Bender et al. 1996, p. 4 et seq.; Wermter

and Cramer 1988). Under this procedure, employers are required to submit notifications to the responsible social security agencies concerning all of their employees covered by social security at least once a year. The BeH covers all white- and blue-collar workers as well as apprentices as long as they are not exempt from social security contributions. This means that civil servants, self-employed persons and regular students[9] (see Cramer 1985) are not recorded in the BeH. After the notification procedure was changed on 1 January 1999, employees in marginal part-time employment and unpaid family workers have also been recorded (not contained in the data until 1 April 1999). The data are recorded by the health insurance companies, collected and edited by the Federal Employment Agency (BA) and subsequently integrated into the History File by the IAB.

The administrative individual data are supplemented with establishment data (Basis Establishment File and BHP Extension Files). They are taken from the Establishment History Panel (Betriebs-Historik-Panel - BHP), which is also based on the BeH (see Schmucker et al. 2016).

When linking individual data with establishment data it has to be taken into account that the variables in both the Basis Establishment File and the BHP Extension Files are measured and aggregated at the reference day 30 June in each year.[10]

The Basis Establishment File is linked with the Individual File via the programme-specific commands of the software packages used for preparing and analyzing the data. In Stata, for instance, the two files can be linked using the "`merge`" command in connection with the relevant paths (see Box 1).

```
* Load INV-SIAB data module
use INV-BIO_8014_v1_INV-SIAB.dta, clear
gen int jahr = year(begepi)

* Merge establishment level information from INV-BHP
sort betnr jahr
merge m:1 betnr jahr using INV-BIO_8014_v1_INV-BHP_Basis_v1.dta
```

Box 1: Example code for Stata

2.2.2 Benefit Recipient History (LeH)

The Benefit Recipient History (Leistungsempfänger-Historik - LeH) of the IAB covers periods during which individuals receive earnings replacement benefits from the Federal Employment

[9] Students may still appear in the BeH if, for example, they have a marginal part-time job parallel to their courses.

[10] An extreme example: an employment notification exists from 1 January 2006 to 30 May 2006; the establishment goes bankrupt in June 2006. There is then no information about this establishment in the BHP for 2006.

Agency (sphere of Social Code Book III). The benefits comprise unemployment benefit, unemployment assistance and maintenance allowance, in other words not benefits under the sphere of Social Code Book II (e.g. unemployment benefit II). Since entitlement to benefits depends on meeting certain legal requirements, periods of unemployment in which the requirements are not met (e.g. no eligibility for unemployment assistance, or non-completion of the qualifying period for unemployment benefit) are not reported in the Benefit Recipient History. The earliest available data in the LeH are from 1 January 1975.

2.2.3 Unemployment Benefit II Recipient History (LHG)

The Unemployment Benefit II Recipient History (Leistungshistorik Grundsicherung - LHG) contains information about individuals who are eligible for benefit and capable of work, about the members of their benefit community (Bedarfsgemeinschaft) in accordance with § 7 SGB II and about certain individuals associated with the benefit community. It is not possible to link individuals with benefit receipt in accordance with Social Code Book II (SGB II) at the level of benefit communities within the INV-SIAB. The receipt of benefits in accordance with SGB II covers both basic social security benefits (e.g. Unemployment Benefit II) and supplements to unemployment benefit or additional benefits. The LHG does not contain any information about the benefit rates, however. As the amount of benefit received is not determined at the level of the individual but at the level of the benefit community in the case of Unemployment Benefit II, it is difficult to assign an individual benefit rate.

Unlike the benefits in the sphere of Social Code Book III, the Federal Employment Agency (BA) is not the sole organization responsible for administering the benefits. The data therefore distinguish between the three possible types organizational responsibility in the legal context of SGB II:

- Cooperation of employment agencies and municipalities (Arbeitsgemeinschaften – ARGE) until the end of 2010 / joint facilities (gemeinsame Einrichtungen) since 2011), in which the BA and the municipality deal with tasks jointly,
- separated responsibilities (getrennte Trägerschaft) / municipalities exercising their duties separately (until 2011) – here the tasks are divided between the BA and the municipality[11],
- authorized municipalities, which are also called opting local authorities or opting municipalities according to the initial experimental clause of Section 6a - here the local authority is responsible for all tasks in the sphere of SGB II.

[11] The municipality pays the costs for housing and heating (Section 22 SGB II) and additional one-off benefit payments to cover extra costs (Section 23 (3) SGB II) and the additional benefits to support integration in accordance with Section 16 (2) Clause 2 No. 1 - 4 SGB II. The BA, on the other hand, covers the costs for regular benefits, social security contributions and integration benefits (SGB III and SGB II) and specific benefits excluding the additional benefits to support integration cited above.

The data of the “Unemployment Benefit II Recipient History drawn from A2LL” (LHG) come from different reporting procedures. As a rule, the IT procedure A2LL was used in all ARGE cooperation projects until 2010, and in joint facilities from 2011 onwards[12]. Authorized municipalities use various IT procedures of their own and transmit their data to the BA by means of the XSozial-BA-SGB II standard. Both procedures are used by municipalities with separated responsibilities. The different data standards affect the scope and quality of the data supplied.

The earliest available data in the LHG are from 1 January 2005. However, the data source is incomplete until the beginning of 2007 (see Section 0).

2.2.4 Jobseeker History (ASU / XASU)

Data about jobseekers are stored in the Jobseeker History (Arbeitsuchendenhistorik – ASU / XASU). The ASU data source contains information on jobseekers who are registered with employment agencies, and from 2005 onwards also includes ARGE cooperation projects and separated responsibilities for the implementation of SGB II. The XASU data source, on the other hand, contains the data of jobseekers in receipt of Unemployment Benefit II (ALG-II) from authorized municipalities from 2005 onwards. These data are reported in accordance with the X-Sozial-BA-SGB II standard.

2.2.5 Participants-In-Measures History Files (MTH)

The Participants-In-Measures History Files (Maßnahmeteilnahmehistoriken - MTH) contain information that can be assigned to different legal spheres. First, they contain active labour market policy measures in accordance with Social Code Book III and participation in such measures. Second, the MTH contain measures in the legal sphere of Social Code Book II if these are recorded in BA administrative procedures. This means in particular that no measures implemented by the authorized municipalities are recorded in the MTH as these are reported via a different standard, XSozial. Information from these institutions is not included in the IEB due to a number of data problems. The earliest available data in the MTH are from 1 January 2000.

2.3 Administrative Data from the IEB

2.3.1 Corrections and validation procedures

Before the data from the data sources specified in Section 0 are merged to form the IEB they undergo source-specific correction procedures (see the following Sections). The IEB as a whole undergo the following corrections:

[12] In 2014 A2LL was gradually replaced by ALLEGRO as the new IT procedure for Unemployment Benefit II in the sphere of SGB II in joint facilities.

- Observations in which the age is under 13 or over 75 are deleted.
- Observations whose end date precedes the start date are deleted.
- Inconsistent information on gender or date of birth within an account is corrected.
- Records with no information on the date of birth or on gender after the correction procedure are deleted.
- No further corrections (such as the addition of likely missing notifications, strike corrections) are implemented.

2.3.2 Employee History (BeH)

- To capture a person group that is as constant as possible over time, some person groups for which data are not available throughout the entire observation period are excluded. From the reporting year 2011 onwards the BeH data originate from newly designed source data. As a result, a number of person groups have been introduced or reactivated as they are classified by the BA statistics as being subject to social security contributions. The person groups 101 - 107, 111 - 114, 118, 119, 120, 140, 141, 142, 143, 149, 201 and 203 - 205 are therefore contained from that time onwards as well as the two groups 109 and 209, which indicate people in marginal part-time employment. Groups that are not included are, for example, people in short-term employment, i.e. person groups 110, 202 and 210.
- Person groups 123, 124 and 127 have been newly introduced.
- For data protection reasons, the person groups 107, 111, 113, 114, 127 and 204 are combined to form the person group “other workers” (599).
- From the reporting year 2012 onwards apprentices were included as the new person groups 121 and 122.
  - Observations with earnings amounting to zero or with no details on earnings, and the value 101 for the person group variable, and the value 50 for the reason for notification (annual notification) are not incorporated into the IEB.
  - Gender and date of birth are taken from the Data Warehouse (DWH) of the BA. This information is harmonized across data sources.
  - The territorial allocations for place of work and place of residence are updated to the status as of 31 December 2016.

2.3.3 Benefit Recipient History (LeH)

- Observations without a valid start date are excluded.
- Observations whose end date precedes the start date are excluded.
- If the end date for the receipt of unemployment assistance precedes the start date by one day and the spell was not deleted, then the end date is increased by one year.

- Between 2004 and 2006 the notification procedure from which the data originate was changed. Overlaps occurring between the old and the new procedures were corrected.
- Observations with no end date or an invalid end date are excluded, since in these cases it cannot be assumed that a benefit payment was made at all.
- The territorial allocations are corrected in the same way as the BeH in the SIAB.

2.3.4 Unemployment Benefit II Recipient History (LHG)

- Observations without a BA client number are deleted.
- Observations without a valid date of birth are deleted.
- Cancelled data records are not used.
- It only contains observations of people who are capable of work and people under the age of 65.
- In each case non-overlapping periods of benefit entitlement of a person in a certain benefit community are depicted. New observations are begun for the following administrative reasons:
  - on certain birthdays of members of the BG that are stipulated by law and relevant for structural changes in the benefit community (14, 15, 18 and 25) and the individual retirement age of members of the BG,
  - when the structure of the benefit community changes (e.g. due to entries/ exits),
  - when there are changes in a variable of the BG client and
  - at the beginning and the end of a case of benefit sanctions for observations from 1 April 2006 onwards. It must be taken into account, however, that it is not possible to identify the duration or type of sanction or the time when it was imposed or when it began on the basis of the data. The reason for this is that there is no corresponding variable or value that indicates the start, type or duration of the sanction.
- For the reason mentioned above, all individual-related variables that are available for the LHG source are valid for the entire duration of the observation.
- Double notifications due to the territorial reforms in 2009/2011 and the reorganization of the institutions in 2012 were corrected as far as possible.
- The territorial allocations are corrected in the same way as the BeH in the INV-SIAB.

2.3.5 Jobseeker History (ASU / XASU)

- Observations whose end date precedes the start date are not included in the ASU.
- There is no consolidation of the ASU observations for individual persons. Therefore, overlaps between ASU observations might occur.
- Individual-related variables that are only available for the (X)ASU sources always refer to the beginning of the spell.

- A new ASU spell is generated as soon as a change of status occurs (e.g. from seeking work to unemployed). This also applies if the type of institution (employment agency, cooperation of employment agency and municipality, joint facility, authorized municipalities, separated responsibilities) changes. The ASU data basis only distinguishes between observations with the status “unemployed” and “jobseeker”, and since 2006 “seeking advice” and “without status”. In the IEB, however, the additional status “ill / not able to work” is available. The employment status “ill / not able to work” is assigned to IEB spells when in the ASU data basis
  - a preceding observation with the status “unemployed” exists which joins the next observation without a gap and has “incapacitated for work” as the reason for exit, and
  - a subsequent observation with the status “unemployed” exists which also follows without a gap, and
  - the observation itself does not have the status “unemployed” but “jobseeker”.
  - In contrast to the ASU source, the XASU only distinguishes between the status
    - “not unemployed, but seeking work” or
    - “unemployed and simultaneously seeking work”.
  - Unlike in the ASU, periods of illness are not taken into account when generating the “employment status”, since no information about illness is available in the XASU data. When calculating the unemployment duration with XASU observations, gaps due to illness cannot be identified.
  - The XASU contains non-overlapping time periods for individuals. If one of the following variables changes, in each case a new data spell is generated for the XASU:
    - change of job search status
    - change of availability
    - change of SGB II institution (due to notification procedure)
    - change of place of residence
  - The territorial allocations are corrected in the same way as the BeH in the INV-SIAB.

2.3.6 Participants-In-Measures History Files (MTH)

- Observations whose end date precedes the start date were excluded.
- Observations generated more than a year after the end of the measure are deleted if another observation exists that was generated within the year after completion of the measure.
- Only the most recent record of an individual case of participation in a measure is used.
- Only cases of participation in measures that are classified as “actually took place” are included in the IEB. Cases of participation that did not take place or have not yet taken

place are deleted. Cases of participation are also classified as not having taken place when a deletion date is set during the participation in a measure.

– Certain types of measure are not included. These include services to support careers advice and job placement, mobility assistance and pure rehabilitation measures.

2.3.7 Episode splitting

The individual labor market biography data are available in the structure of “split” episodes. If episodes overlap each other, these observations are replaced by artificial records with new dates so that completely parallel periods and non-overlapping periods are created. This splitting procedure of otherwise overlapping episodes increases the number of records in the data set (see Figure 4).

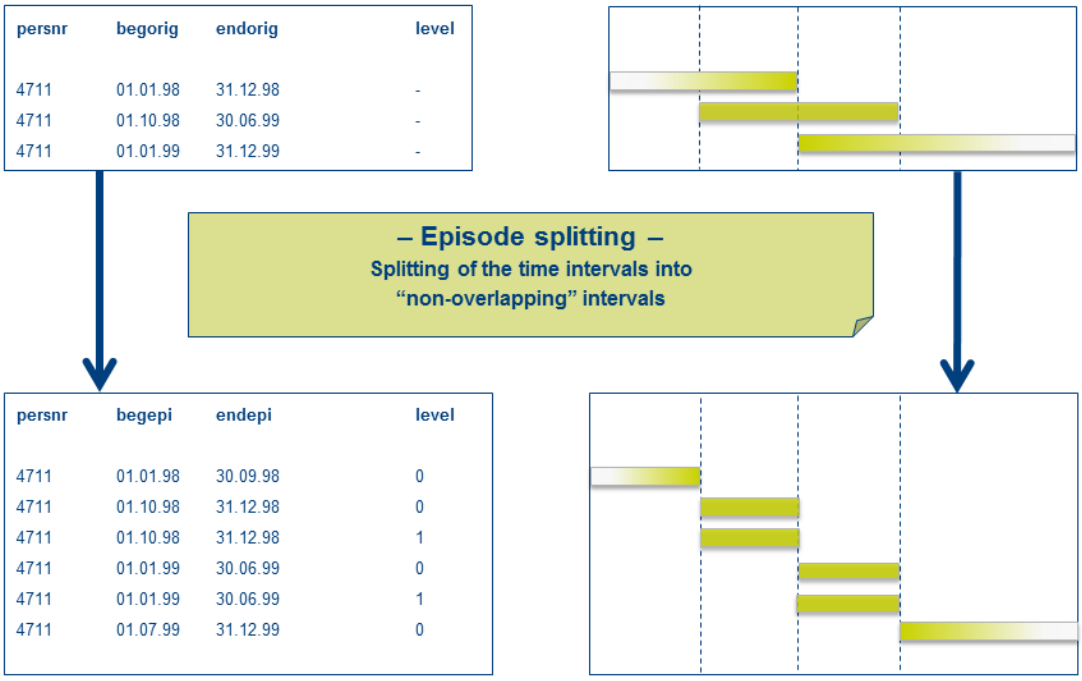


Figure 4: Episode splitting

The original date variables for the beginning and the end of the original record (begorig and endorig) are retained, the newly generated variables 'start date of the split episode' and 'end date of the split episode' (begepi and endepi) mark the beginning and the end of the split episodes. It is possible to identify records that were subject to splitting by comparing the original date information (begorig and endorig) with the dates pertaining to the episodes (begepi and endepi).

To restore the original data without the split episodes or to delete the episodes that were created artificially by means of episode splitting, it is necessary to select all observations for which the start of the original observation is the same as the start of the split episode

```
*  Restore  original  data  structure  of  unsplitted  episodic  data
count
keep if begepi == begorig
drop begepi endepi
count
```

Box 2: Example code for Stata

To facilitate data processing, it is advisable to sort parallel episodes generated by the splitting procedure in a consistent manner. The variables 'observation counter per episode' (level2) and 'observation counter per episode and source' (level1) that can be generated using the following Stata commands, if required:

```
* Counter for parallel episodes:
bysort persnr begepi quelle (spell): generate long level1 = _n-1
label variable level1 "observation counter per episode"

* Counter for parallel episodes in the same source (e.g. identifies
parallel employment episodes):
bysort persnr begepi (spell): generate long level2 = _n-1
label variable level2 "observation counter per episode and source"
```

Box 3: Example code for Stata

2.4 Sampling procedure

The INV-BIO data represent the linked population of inventors who were identified as employees from 1999 to 2011 with their full biographies. A more detailed evaluation is presented in the Appendix.

2.5 Missing values

In the data, missing values are coded as follows (see Table 5):

Term	Value	Description
No (valid) details available	.z	Values of a variable which are not systematically missing, i. e. the variable is available in principle for the data source, but no details are available for the value considered or cannot be interpreted reasonably.
Systematically not available	.n	A variable is by definition not available for a data source (dark grey cells in the overview of variables in Section 0) or is not defined for a certain period.

Table 5: Coding of missing values

3. Data quality and problems

3.1 Patent register data and inventor patent file

- Patent information in the INV-BIO is restricted to the period ranging from 1980 to 2014.
- The INV-PAT file is structured by inventors and DOCDB patent families. Patent families represent a group of patent applications, which refer to the same technical content, but are split into multiple patent applications (EPO 2017). DOCDB patent families are defined in the process of patent examination. For a more detailed discussion of patent family concepts, see Martinez (2010). Analyzing patent families in contrast to patent applications reduces potential bias introduced by divisional patents, i.e. patents that are across different jurisdictions (here: EP and DE). Figure 5 visualizes the change in the data structure of the inventor-patent track records that are performed towards the generation of the INV-PAT data set.

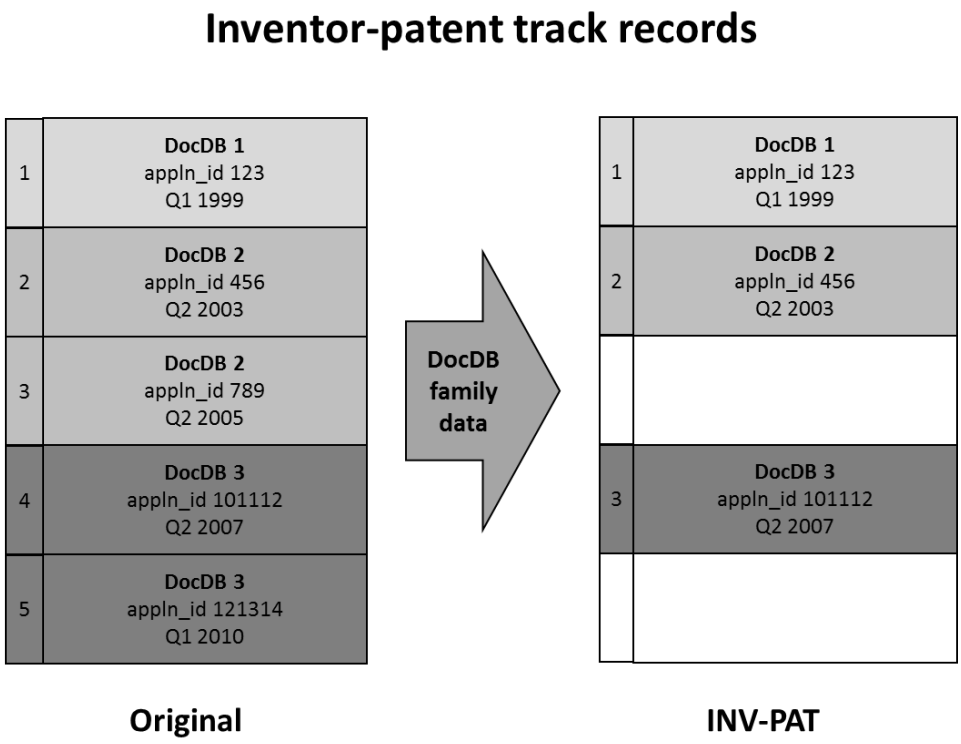


Figure 5: Transformation of inventor-patent track records for INV-PAT

- All measures take patent filings into account irrespective of whether the patent has (already) been granted or not. All patent-related information comes from the PATSTAT database (version 2017-10).
- All dates in the patent register data are recorded at a daily basis for all legal events (e.g. patent filing, patent publication, and grant) in the life cycle of a patent. In the INV-BIO

research data set, the dates are converted to a quarterly structure for data privacy reasons. The date variables included in the INV-BIO data set refer to the first patent application within the DOCDB patent family that was filed by the focal inventor and that was assigned to the unique inventor in the matching procedure (see Figure 5). This reference application does not correspond to the earliest patent filing in the DOCDB family in about 20% of the cases. An analysis of the dates of the reference filings shows that 99 percent of the deviations are within a time window of one year (366 days) from the earliest filing within the DOCDB family – the so-called priority year.

- Patent examiners assign technology codes in the International Patent Classification system (IPC) to each patent application. To determine the technology classification of patents, all individual IPC classes assigned to a patent were reclassified using the WIPO technology concordance table (Schmoch 2008) into a classification of technology areas. Departing from this list of technology areas, the modal technology field was determined as representative for the patented invention. The technology main area, a coarsened version of this technology classification system with 5 classes, is available by default in the INV-BIO data while the detailed classification of 34 technology areas is only available on request. The technology information used to characterize the patent family is drawn from the reference application within the family (see Figure 5). At the patent family level, about 2% of the inventions differ in their technology information (TF34 main area) in the INV-PAT data.
- Forward citation variables in the INV-BIO data use the first publication date of a patent application within the family as the reference date to construct time windows in which citation counts for different jurisdictions (DE, EP, US) were generated. Patent applications in the later years might be subject to censoring of the forward citations. For instance, our version of PATSTAT (version 10-2017) gives only proper forward citation counts until the 2017. Hence, patent filings recorded in 2014 will subject to censoring after 3 years and citation windows exceeding this margin will be redundant. We advise researchers to take censoring of the forward citations into account in their research.
- For data privacy reasons, the number of forward citations, which are heavily skewed, is truncated at the 99th percentile. Patents that received more citations than the 99th percentile are recorded in the data with the truncation threshold value. The cutoff is implemented individually for each variable.
- The numbers of inventors and applicants as well as the number of foreign inventors or applicants are based on the (original) address information recorded in the patent register (person_country_code) for the two groups. The information provided in the INV-BIO data refers to the most recent patent publication.

- Applicant information may vary over time as a result of e.g. ownership changes of the applicant or transactions of patents. We document the applicant count as recorded on the priority application within each patent family. In contrast to applicants, inventor information remains very stable as events that cause a change in the inventor are very rare. For further information on patent transfers in Europe, see Gaessler (2016).
- Information on granted patents and the grant date are only made available for the reference application of the patent family. An indicator variable marks patent families that have one or more granted patents and the earliest grant date at the level of the DOCDB family.
- Average distances in kilometers computed between the inventors' workplaces as listed on each patent (family) are based on detailed geographical data at the municipality level that are available in the social security register. For data privacy reasons, municipality level data in the INV-SIAB are not available to external researchers.

3.2 Entire IEB

The IEB contains employment histories. However, not every type of employment is included in the administrative data. Some individuals with certain life courses are not represented in the IEB at all.

For evaluation purposes, it is often relevant to know gaps in the included biographies (e.g. for creating control groups, analyzing life courses, etc.). The gaps listed below are defined as periods of time after the end of school education for which no data are included in the IEB. These gaps can be divided into

- gaps with no information available at all, and
- gaps for which information may be available from the 'reason for notification / reason for end of benefit receipt / reason for discontinuation of SGB II / reason for deregistration' variable of the observation immediately preceding the gap (if a corresponding observation exists).

These gaps were identified using the 'reason for notification / reason for end of benefit receipt / reason for discontinuation of SGB II / reason for deregistration' and 'employment status' variables in the various sources. The list makes no claims to be exhaustive.

Biographical gap	**Information on gap, identifiable using the details in the "grund" variable in the preceding observation of the source, if necessary**
Civil servants, professional soldiers, judges, employees of bodies or foundations under public law	XASU

Self-employed persons without support	LeH, ASU
Students, persons in school-based further education	LeH, LHG, ASU, XASU
Persons who are ill / not able to work for more than 6 weeks (illness during unemployment, however, is represented in the ASU source under certain circumstances, see Section 2.3.5)	BeH, LeH, ASU
Persons receiving old-age pension without employment if not a member of a benefit community	LeH, LHG, ASU
Individuals on maternity leave / parental leave	XASU
Recipients of early retirement benefits	LeH, ASU
Trade professionals working from home	
Employees working short-time	ASU
Persons in youth welfare facilities, in vocational training centres, approved workshops or similar facilities for disabled persons	ASU
Participants in programmes to support participation in working life (people in rehabilitation)	ASU
(Sideline) farmers	
Caregivers according to Section 19 SBG XI	
Conscripts	BeH, LeH, LHG, ASU, XASU
Persons in reserve duty training	BeH, LeH, LHG, ASU, XASU
Persons fulfilling community service	BeH, LeH, LHG, ASU, XASU
Persons fulfilling a voluntary social or ecological year instead of community service	
Other people not registered with the statutory pension insurance or the Federal Employment Agency (e.g. sabbatical, funding from personal assets or pensions, emigration, employment abroad, voluntary work etc.)	BeH, LeH, ASU
Strikers in cases where the strike lasts more than a month	LeH
Social assistance recipients (prior to the introduction of SGB II in 2005), recipients of welfare payments (according to SGB II)	
Recipients of compensation according to FELEG (Gesetz zur Förderung der Einstellung der landwirtschaftlichen Erwerbstätigkeit, Act on Support in Case of Termination of Farming Activities)	

Table 6: Biographical gaps and identification possibilities

3.3 Employee History (BeH)

- For consistency with the patent data, employee history information in the INV-BIO data is restricted to the period ranging from 1980 to 2014.
- The introduction of the occupational classification KldB 2010 in 2011 led to a number of problems. For example, during the transition period granted to employers in the social security notification procedure,[13] there was a temporary increase in the number of missing details. Analyses of the BA statistics (Bertat et al., 2013, p. 10) show that in 20 to 30 percent of cases no information was contained in the new or converted variables 'occupation – activity performed', 'working hours' and 'vocational education and training' after the switchover. This situation began to improve significantly in the first half of 2013. In order to improve the quality of the 'working time' variable in the transition period, Ludsteck and Thomsen (2016) developed an imputation procedure to replace the missing values by imputed values. The imputed variables are already used in the INV-BIO data.
- Due to the introduction of the employment notification procedure in the federal states of East Germany, the notifications for East Germany can only be assumed sufficiently complete from 1993 onwards. For the same reason, a large number of episodes in 1991 have missing values for several variables (e.g. employment status).
- The increase in the number of BeH observations from 1999 onwards is due to the introduction of the obligation to submit employment notifications for people in marginal part-time employment from 1 April 1999 onwards.
- Especially in 1999, observations of part-time employment increase significantly. This is caused by the actually observed increase in part-time work as well as by the fact that since 1999 employment notifications have been completed more correctly.
- Within the employment notification procedure, a certain time lag is unavoidable. Although changes in employment relationships have to be reported immediately and existing employment relationships have to be confirmed annually by April (until the annual notification 2012) or mid-February (from the annual notification 2013 onwards) of the following year, some notifications actually arrive years later. The History File of the IAB is not updated continuously, however, but at certain intervals. This is done using files of employment notifications for one particular year which were submitted 36, 18, 12 or 6 months after the end of the reporting year (e.g. the 18-month file for 2013 can be created in July 2015 at the earliest). Notifications submitted more than three years late are not

[13] The test programs used in the notification procedure permitted missing details in the occupation code 2010 until the end of May 2012.

taken into account at the IAB, which means that a 36-month file shows a 100 % degree of completeness by definition.

- In the version of the IEB on which the INV-BIO ADIAB data are based, the degree of completeness of the BeH observations is 100 percent for the year 2013. Information censored at 18-months were used for 2013 and 2014. Hence, BeH observations for 2013 and 2014 may be missing for some establishments as they might not have submitted their employment reports within this 18 month period. Missing data leads to gaps in the employment data, which, however, will be replaced with actual reports in future register extracts that include the delayed records. Missing information that is due to this feature of the register is clustered by establishments, hence establishment information may be incorrect and is subject to changes in future version of the data.
- In 1984, the employment notification procedure was subject to a change. From that time onwards, one-off payments of gross earned income were reported as part of the annual earnings subject to social security contributions, which leads to an increase in the average daily wage. In particular, the proportion of wages and salaries above the upper earnings limit increased considerably from that year onwards (see Bender et al. 1996).
- For the years 1992 until 2000 noticeable decreases and increases in the number of notifications were observed. Decreases can be observed especially for the following 10 districts: Braunschweig (03101), Salzgitter (03102), Wolfsburg (03103), Emden (03402), Kassel (06633), Essen (05113), Neuss (05162), Erftkreis (05362), Hersfeld-Rotenburg (06632), Miltenberg (09676), Kempten (Allgäu) (09763), Hoyerswerda (14264). This is due to notification problems of one or more establishments located in these regions.
- Concerning the notifications for full-time employment, especially the districts Main-Taunus (06436) and Alzey-Worms (07331) are noteworthy. They feature above-average increases of the number of full time employees. Also, the reasons are notification problems at one or more establishments located in these regions.
- In the years 1996 to 1998, the values 841-844 (doctors and pharmacies) within the ‘occupation – activity performed’ variable are very rare compared to other years.

3.4 Benefit Recipient History (LeH)

- For consistency with the patent data, the benefit recipient history information in the INV-BIO is restricted to the period ranging from 1980 to 2014.
- For the states of eastern Germany, the LeH observations were not fully recorded until 1992.

- The benefit receipt data used to be saved on magnetic tapes. Owing to a fault in one magnetic tape, the benefit receipt data up in 1980 are only partially contained and incomplete.
- Due to an internal change of systems, there is a break in the recording of periods of exclusion from benefits and of benefit suspension in 2004. Until 1 July 2004 periods of exclusion from benefits and of benefit suspension can only be identified via the 'reason for end of benefit receipt' in the preceding LeH observation. After this date a separate observation is available with the daily benefit rate = 0 for periods of benefit exclusion and suspension.

3.5 Unemployment Benefit II Recipient History (LHG)

- With regard to the completeness of case numbers or benefit histories from the LHG data sources, there are substantial gaps in the years 2005 and 2006. We therefore strongly advise against analyzing the data for this time period based merely on the LHG sources.
- Longitudinal analyses of individuals are affected by inaccuracies as it is not possible to distinguish between changes in the benefit entitlement status and relocations into and out of districts whose institutions had problems delivering data.
- Also from 2007 onwards, cases of under reporting occur at times. These generally last one month and occur mainly in the authorized municipalities.
- Disproportionate reporting bias occurs with changes in the type of institution responsible for implementing SGB II:
  - In the context of the reform of the territories covered by the institutions, which came into force on 1 January 2011, cases of underreporting occurred in the districts covered by the employment agencies of Dessau-Roßlau, Halberstadt, Halle and Sangerhausen.
  - Double notifications due to the territorial reforms in 2009/2011 and the changes in the form of the institutions as of 1 January 2012 are already corrected as far as possible in the IEB. Nonetheless double notifications may still occur.
- In the following job centers there are inaccuracies with regard to the allocation of benefit cases:
  - between Emden and Norden between September and December 2009
  - between Döbeln and Mittelsachsen from October to December 2012
  - between Tirschenreuth and Wunsiedel from November 2012 to March 2013
- Some individuals for whom a (X)LHG spell exists are excluded entirely or partly from benefit receipt according to SGB II, for instance because they take part in a subsidized training programme, receive an old-age pension, live in an in-patient facility or a residential

institution or receive insurance payments aimed at avoiding need. This affects on average 3 to 5 percent of all cases. In XSozial this person group is sometimes underreported by some institutions. Exclusion from benefits cannot be identified in the INV-SIAB.

3.6 Jobseeker History (ASU / XASU)

3.6.1 ASU

- The registered periods of job search activity in the ASU source are regarded as complete from the year 1997 onwards. Therefore, the analysis potential of the ASU spells before 1997 is limited.
- For the placement staff it is not always possible to record the allocation to the legal sphere immediately, since it is frequently only clear which institution is primarily responsible after a certain time due to a possible entitlement to SGB II benefits. Therefore, we recommend comparing the value of the 'type of institution' variable in the ASU with the value in the LHG and/or XLHG for the same period of time. Due to the recording gaps in the LHG and XLHG between 2005 and 2006 this is not always possible.
- From mid-2005 until mid-2006, the coArb IT procedure, from which the jobseeker and applicant pool data originate, was superseded by the VerBIS procedure at the Federal Employment Agency. In July 2005, coArb was first replaced by VerBIS in the employment agency in Wiesbaden as a pilot project. From December 2005 onwards, it was then gradually replaced by VerBIS in several stages in all employment agencies. The information for many of the variables recorded was gathered with different levels of differentiation and different qualitative weighting in the two systems. It is therefore very difficult to integrate these variables into the IEB, which is only possible using a special procedure (mapping). Unfortunately, a full conversion of the affected variables from coArb to VerBIS cannot be achieved by means of mapping, so for some variables there is an unusually large number of the values 'no details available', 'other' or 'missing'. Moreover, striking differences may occur in frequency counts, depending on whether the original source of the data was coArb or VerBIS. Important limitations in the analysis potential are mentioned in the corresponding description of variables.
- The coArb procedure, which was used until June 2006, supported only the placement of unemployed persons and jobseekers. Some data were also collected about individuals who were only seeking advice, but these data are incomplete. The careers advice data were collected in a separate system. In VerBIS the attributes of the job-search status were extended to include 'seeking advice' and individuals 'without status'. The latter group includes individuals eligible for Unemployment Benefit II who are only available for job

placement to a limited degree. The recording of this group in VerBIS is only regarded as largely complete since January 2008.

- A change of the institution responsible for implementing SGB II or a change of place of residence does not lead to a new ASU observation, the value of the variable at the start of an episode is continued. The longer the observation becomes, the greater the risk is that the institution responsible or the place of residence is no longer correct.

3.6.2 XASU

- In contrast to the job search spells from the cooperation of employment agencies and municipalities (ARGE) and the separated responsibilities, systematic cases of under reporting are documented for the authorized municipalities since 1 January 2005. Thus, data from the XASU source should only be analyzed from 2007 onwards.
- A variety of variables sometimes have only a very low degree of completeness for the XASU. Variables which are affected by this include ‘school-leaving qualification’, ‘severe disability status’, ‘reason for notification’ as well as ‘employment status prior to job search’. Although the degree of completeness of these variables improves over time, some of them are still unsatisfactory. The ‘occupation – activity performed’ variable is not available in the XASU for almost the entire period available.
- For a number of institutions (districts), the proportion of registered recipients of unemployment benefit II who are also registered jobseekers is implausibly large at times or continuously in the IEB. One possible reason for this could be an incorrect determination of the status ‘not unemployed but seeking work’ by these institutions.
- The institution-related and period-related plausibility of the XASU data should be examined before use, taking the research question into account.

3.7 Participants-In-Measures History Files (MTH)

- The MTH is incomplete for measures with a start date earlier than 1 January 2000.
- As of 1 January 2005 there is an inconsistency in the data as participants in measures were allocated to different institutions with the introduction of Social Code Book II (see Sections 2.2.2 and 2.2.4).
- The MTH contains only notifications that are recorded in BA procedures. The use of these procedures in cooperations of employment agencies and municipalities/separated responsibilities/municipalities exercising their duties separately increases continuously between 2005 and 2007. The notifications for these institutions are complete from March 2007 onwards.

- Measures that are reported via the XSozial standard are not contained in the MTH or in the INV-SIAB.
- As a result of the reorganization of the institutions responsible for implementing SGB-II, the documentation of participation in measures in the MTH may end or begin again when there is a change in the reporting procedure.[14]

4. Description of variables

Frequency counts and overviews of the individual values and labels of the variables can be found in separate files under http://fdz.iab.de/en.aspx.

4.1 Identifiers

4.1.1 Inventor ID (erf_id)

Variable label	Inventor ID
Variable name	erf_id
Category	Identifiers
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The pseudonymized individual ID groups observations by unique persons. It is not possible to infer any of the person's characteristics or any original identifiers from this individual ID. The actual formation of the individual identifier, which spans all data sources (e.g. employment and benefits) is based on a heuristic that was developed by the BA. erf_id groups inventor-patent records in the INV-PAT data set by individuals. The link between patent records and individuals in the IAB data was established using a methodology combining record linkage techniques and machine learning. For further details, see Appendix A2.

4.1.2 Establishment ID (betnr)

Variable label	Establishment ID
Variable name	betnr
Category	Identifiers
Origin	BeH
Data type	Numerical
Hierarchy	None

[14] Further information on the territorial structure of the institutions responsible for implementing Social Code Book II is provided by the BA at http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Regionale-Gliederungen/Gebietsstruktur-Traeger-Grundsicherung-Nav.html

Detailed description	The establishment ID indicates which observations belong to the same establishment. It is based on the establishment number allocated by the BA, which was replaced by an artificial number. Further information on the allocation of establishment numbers by the BA can be found in Bender et al. 1996: p. 15 et seq. and pp. 27-30. The establishment number and year specification can be used to merge individual and establishment information. For the establishment number, the following should be observed in general: a) If the company has one office only, or if the company has one office only in one municipality, this office is the establishment and is given an establishment number. b) If the company has **several** branch offices in **one** municipality, these establishment premises / workplaces must be merged into a single establishment under one establishment number, if they belong to the same economic class. If they do **not** belong to the same economic class, each branch office is regarded as a separate establishment and is given its own establishment number. c) If the company has **several** branch offices in **several** municipalities, each of these branch offices is an establishment and is given its own establishment number. In this context, the following definitions with regards to the allocation of establishment numbers as part of the notification procedure for social security must be observed: a) An **establishment** is a regionally and economically delimited unit in which employees work and which is allocated an establishment number according to the above-mentioned principles. b) A **workplace** is a unit in which employees work and which is not allocated an establishment number according to the above-mentioned principles. c) A **company** as a term combines establishment premises and workplaces belonging to the same employer. d) An **employer** is any natural person or legal entity that employs at least one employee subject to social security contributions or in marginal part-time employment. e) Establishment and establishment premises are synonyms; branch office is a synonym for subsidiary, district office, outsourced office, workplace etc. if it is not an establishment.
Notes on quality	The establishment ID is only missing in a very small number of cases. These observations are notifications for the person group “205” (earnings notifications for casual workers). As establishment variables (place of work, economic activity, establishment size etc.) are merged via the establishment ID, they are missing in these observations.

4.2 Generated technical variables

4.2.1 Observation counter per person (spell)

Variable label	Observation counter per person
Variable name	spell
Category	generated technical variables
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The observation counter per person counts a person's observations, beginning with 1 to N. The variable is generated during the episode splitting procedure and refers to the split observations. Using the "observation counter per person" variable, it is easy to restore the original sorting order. The observations are sorted first by the start date of the split episode and then by the data source.

4.2.2 Data source of record (quelle)

Variable label	Data source of record
Variable name	quelle
Category	generated technical variables
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The variable indicates the data source of the record.

4.2.3 Year (jahr)

Variable label	Year
Variable name	jahr
Category	generated technical variables
Origin	BeH
Data type	Numerical
Hierarchy	None
Detailed description	This variable is only included in the Establishment File. It indicates the year of validity of the establishment data as of the reference date of 30 June. This variable can be used in combination with the establishment ID (betnr) to merge the INV-SIAB and the complementary establishment records.

4.3 Period of validity

4.3.1 Original start date of observation (begorig)

Variable label	Original start date
Variable name	begorig
Category	period of validity

Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Date
Hierarchy	None
Detailed description	The original start date of the observation corresponds to the original start date of the notification. This can differ from the start date of the episodes (begepi) (see also the comments on episode splitting in Section 2.3.7) **1) BeH** Because of the rules of the notification procedure, for BeH observations the starting and ending year are always identical (obligation of the employer to submit annual employment notifications). A continuous employment relationship may therefore be distributed across several notifications. **2) LHG, ASU, XASU** begorig indicates the start date of the new period.

4.3.2 Original end date of observation (endorig)

Variable label	Original end date
Variable name	endorig
Category	period of validity
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Date
Hierarchy	None
Detailed description	The original end date of the observation corresponds to the original end date of the notification. This can differ from the end date of the relevant line of data, the so-called end date of the split episode (see also the comments on episode splitting in Section 2.3.7). **1) BeH** Because of the rules of the notification procedure, in BeH observations the starting and ending year are always identical (obligation of the employer to submit annual employment notifications). A continuous employment relationship may therefore be distributed across several notifications. **2) LHG, ASU, XASU** endorig indicates the end date of the new period.

4.3.3 Start date of split episode (begepi)

Variable label	Episode start date
Variable name	begepi
Category	generated period of validity
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Date

Hierarchy	None
Detailed description	The start date of the split episode is always equal to or greater than the start date of the original observation (see also the comments on episode splitting in Section 2.3.7).

4.3.4 End date of split episode (endepi)

Variable label	Episode end date
Variable name	endepi
Category	generated period of validity
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Date
Hierarchy	None
Detailed description	The end date of the split episode is always equal to or smaller than the end date of the original observation (see also the comments on episode splitting in Section 2.3.7).

4.4 Personal information

4.4.1 Gender (frau)

Variable label	Gender
Variable name	frau
Category	personal variable
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	Indicator variable for Gende: 0 – male, 1 – female. By definition, gender does not change within individual accounts.

4.4.2 Year of birth (gebjahr)

Variable label	Year of birth
Variable name	gebjahr
Category	personal variables
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The year of birth is constant within individual accounts.
Notes on quality	In the original data, it is possible that erroneous person links across the register data sources cause inconsistencies of the date of birth (and the year). The data preparation corrects these errors by prioritizing the date of birth information from the social security ID number over data that originate in registers of the Federal Employment Service.

4.4.3 Nationality (nation)

Variable label	Nationality
Variable name	nation

Category	personal variables
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The variable contains the nation codes used by the Federal Statistical Office (Statistisches Bundesamt).
Notes on quality	Due to the sensitivity of the information on nationality with regard to the data protection legislation, this variable is only made available in full detail on application and only in well-founded cases.

4.4.4 Nationality, aggregated (nation_gr)

Variable label	Nationality, grouped
Variable name	nation_gr
Category	personal variables
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The variable contains a grouped version of the nation codes used by the Federal Statistical Office.

4.4.5 Marital status (famst)

Variable label	Marital status
Variable name	famst
Category	personal variables
Origin	LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	This variable describes the marital status. The characteristic in the LeH source has only two values (0 - not married, 1 - married), while in the LHG/ASU/XASU/MTH sources, a distinction is made between six values. Inconsistencies across the data sources are possible. The variable included in the data is unedited and no harmonization of the information across the source databases was implemented.

4.4.6 Number of children (kind)

Variable label	Number of children
Variable name	kind
Category	personal variables
Origin	LeH, LHG, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None

Detailed description	The content of the variable differs across the data sources. **1) LeH** Indicates the number of children with age below 16 years at the time when the application for benefits was filed. The variable is an indicator that documents only individuals with - 0 - without children - 100 - one or more children. The variable is not updated regardless of changes in the type of benefit or the approval of benefits. Updates occur only only after an episode of employment which leads to the termination of benefit approval. Therefore, births of children during episodes of benefit receipt are not immediately registered but only in a subsequent LeH episode (if there is any). **2) ASU, MTH** The variable recods the actual number of children. Until 30 June 2006, only up to nine children are recorded. The value 0 is not defined. For observations prior to 30 June 2006, the zero value was recoded to "missing", since it is not clear whether zero should be interpreted as "no children" or as "field not filled in". For observations after 30 June 2006, the variable records only valid information if children actually exist. **3) XASU, LHG** The variable indicates the number of children with age below 16 years who live in a benefit community (Bedarfsgemeinschaft). In the LHG data source, the value is valid for the entire original period.

4.4.7 Vocational training (ausbildung)

Variable label	Vocational training
Variable name	ausbildung
Category	personal variables
Origin	BeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	The content of the variable differs across the data sources. **1) BeH** In the BeH source, the variable records the school education and vocational training as reported by the employers in the employment notification procedure. The following values are defined: 1 No vocational training 2 In-company voc. training/traineeship/external voc. training 11 Degree from a university of applied sciences 12 University degree

	In notifications as of 2011 that record the new occupation code in the KldB2010 system, it is no longer possible to identify graduates of universities of applied sciences, as the new occupation code no longer has a separate category for this qualification level. Graduates from these higher education institutions are recorded with the value 12 University degree. “Changes in the vocational training status frequently occur at the same time as a change of establishment. This is because the notification data are compiled anew in the new firm. If, for example, an employee has gained a higher qualification via a part-time further training course while still working then this change of status is probably not recorded until he/she joins a new firm. It can generally be assumed that when a person is employed in a firm for a longer period, the personal data that they reported when they joined the firm is simply continued.” (Meinken and Koch 2004, p. 63). Methods for correcting missing values or temporal inconsistencies in the education and training data are described in Fitzenberger et al. (2006) and in Drews (2006). The methodology described therein only considers information recorded in the BeH data source. **2) ASU, XASU, MTH** School and vocational education recorded in the sources ASU, XASU, MTH describe the most recent information and highest level of qualification that is documented in the databases. The following values exist: 1 no completed vocational training 2 in-firm vocational training/external vocational training 3 full-time vocational school (Berufsfachschule) 4 technical college (Fachschule) 5 university of applied sciences (Fachhochschule) 6 university 7 vocational education/training not recognized in Germany 8 university degree not recognized in Germany Values 7 and 8 are only valid for the MTH data source.
Notes on quality	The number of missing values in the variable increases continuously over time. In the current version of the data, more than 40% of values are reported with non-valid information (missing). In particular, missing values occur frequently for the employment status groups listed below: - persons in marginal part-time employment, - persons working part-time, - foreign employees and - workers in East Germany.

	A potential explanation for the large and increasing proportion of missing data is that the variable records information that is not mandatory to be reported because it is not required for the calculations of claims in the social security system (see Meinken and Koch, 2004, p. 63). As a result of the migration of the data from coArb to VerBIS system, it is not possible to distinguish correctly between “no completed vocational training” and “no information available” in the ASU and MTH data sources between 2006 and 2008. Thus, if missing data is recorded in this time period it cannot be distinguished between “no vocational training” or “missing data” on vocational education and training. Further, the reason for missing data is unclear as also issues with the classification of education or vocational qualification can lead to missing data. The degree of completeness in the XASU data source is generally low.

4.4.8 School leaving qualification (schule)

Variable label	School leaving qualification
Variable name	schule
Category	personal variables
Origin	BeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	This variable contains the school leaving qualification. The content of the variable differs across the data sources. **1) BeH** Recorded values change with the implementation of the new occupational classification system KldB 2010 in 2011. In the KldB 1988 system, the valid values are: 5 Grade-/lower secondary school with or without leaving certificate, intermediate school leaving certificate or equivalent qualification 8 Upper secondary school leaving certificate from a specialized upper secondary school (Fachoberschule), general upper secondary school leaving certificate, A-level equivalent, qualification for university 9 General upper secondary school leaving certificate, A-level equivalent, qualification for university In the KldB 2010, the variable takes the following values: 1 No school leaving certificate 4 Lower secondary school certificate/ grade school certificate 6 Intermediate school leaving certificate 8 Upper secondary school leaving certificate from a specialized upper secondary school/general upper secondary school leaving certificate, A-level equivalent, qualification for university

	2) ASU, XASU, MTH The following values are defined for these data sources: 1 No school leaving certificate 4 Lower secondary school certificate/ grade school certificate 6 Intermediate school leaving certificate 7 Upper secondary school leaving certificate from a specialized upper secondary school (Fachoberschule) 9 General upper secondary school leaving certificate, A-level equivalent, qualification for university The variable records the information on education at the beginning of the period of job-search or participation in a measure. In the case of people seeking an apprenticeship position, the variable may also contain the school qualification they are working towards in the XASU data source.
Notes on quality	The degree of completeness has been decreasing continuously over time in the BeH to a level of about 66% in the most recent years. In the XASU, it has been increasing continuously and reached a level of about 66% since 2012. In ASU and MTH the coverage is good for the whole time period and without variation over time.

4.5 Information on employment, benefit receipt and job search

4.5.1 Daily wage, daily benefit rate (tentgelt)

Variable label	Daily wage/daily benefit
Variable name	tentgelt
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH
Data type	Numerical
Hierarchy	None
Detailed description	**1) BeH** In BeH observations, this variable records the employee's gross daily wage. It is calculated from the fixed-period wages reported by the employer and the duration of the (unsplit) original notification period in calendar days. The daily wage is documented in Euros. Until 1998, employers in principle only reported the earnings which were subject to social security contributions. Earnings below the marginal part-time income threshold were not reported. Earnings exceeding the upper earnings limit for statutory pension insurance are only reported up to this limit. There are two upper earnings limits in the statutory pension insurance scheme. The earnings limit of the miners' pension insurance is generally higher than the earnings limit of the pension insurance for wage and salary earners. However, it is not possible to differentiate between these two pension systems in the data. Since the inclusion of marginal part-time employees in the employment notification procedure on 1 April 1999, earnings below the marginal part-

	time income threshold have also been recorded; the upper earnings limit still applies as the upper ceiling. In some cases, however, the reported earnings nonetheless exceed the upper earnings limit. Generally, this can probably be attributed to the payment of annual bonuses, which the employer can add to the regular earnings in the annual, employment interruption or end of employment notifications. In this case, it is irrelevant whether the upper earnings limit in the statutory pension insurance, which is decisive for the notification period, is exceeded as a result of this addition. However, such earnings notifications could also be due to incorrect details in the employment period. (The earnings information, however, may be considered less error-prone due to its importance for the calculation of claims.)The marginal part-time income threshold and the upper earnings limit for statutory pension insurance differ from year to year as well as between Eastern and Western Germany (in accordance the location of the establishment). An overview of these limits and thresholds can be found under http://fdz.iab.de. A daily wage of 0 Euros can be put down to “employment interruption notifications”. During these periods, the employment relationship continues to exist in legal terms, but without pay. This is the case for periods of sickness after the end of continued payment of wages, for periods of maternity leave and for sabbaticals. The daily wage is reported with two decimal places. All values greater than 0 and smaller than 0.01 were rounded up to 0.01. This makes it possible to identify the above-mentioned employment interruption notifications with the condition daily wage = 0. **2) LeH** For LeH observations, the variable records the daily benefit rate, converted into Euros in each case. It must be taken into account that for observations with an original start date prior to 1 January 1998, the daily benefit rate is documented for working days, while for records starting from 1 January 1998 (original date) onwards, benefit payments are computed at the level of calendar days. Since 1 January 2005, a daily benefit rate reported as 0 Euros can be put down to benefit suspension periods or interruptions of benefit payments. If a reason for end of benefit is reported for an observation with a daily benefit rate equal to 0, then it is a notification of interruption of benefit payments. In the case of observations that reflect a period of benefit suspension, the entitlement is the same as before the start of the benefit suspension period.

4.5.2 KldB 1988, Occupation main group – current/most recent (beruf1988_2)

Variable label	KldB 1988, Occupation main group – current/most recent, 2-dig. level
Variable name	beruf1988_2
Category	information on employment, benefit receipt and job search
Origin	BeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None

Detailed description	**1) BeH** The occupational title of the job performed by the employee during the notification period is a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes applies to one employee, the employer is required to select the job title that best defines the main activity performed (see BA 2005, p. V). To this end, the employer encodes the employee's job in accordance with the "Classification of Occupations. Systematic and Alphabetical Directory of Job Titles" (published by the Federal Employment Agency, Nuremberg, 1988), which describes approx. 25,000 job titles. The occupational classification is structured in a system 3-digit codes and comprises about 330 values. Employment notifications with an end date later than 30 November 2011 are reported using the new occupational classification 2010 (KldB2010). These values are recoded to the KldB1988 by transferring the key area. This results in inaccuracies. **2) ASU, XASU, MTH** The variable documents the occupation of the last job. See 1) for the determination of the occupation.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure. In the XASU source, the occupation variable is not reported for almost the entire period available.
Anonymization	For data privacy reasons, the 3-digit level of the occupational classification (beruf1988_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the occupational classification (beruf1988_2).

4.5.3 KldB 1988, Occupation group – current/most recent (beruf1988_3)

Variable label	KldB 1988, Occupation group – current/most recent, 3-dig. level
Variable name	beruf1988_3
Category	information on employment, benefit receipt and job search
Origin	BeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	**1) BeH** The occupational title of the job performed by the employee during the notification period is a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes applies to one employee, the employer is required to select the job title that best defines the main activity performed (see BA 2005, p. V). To this end, the employer encodes the employee's job in accordance with the "Classification of Occupations. Systematic and Alphabetical Directory of Job Titles" (published by the Federal Employment Agency, Nuremberg, 1988), which describes approx. 25,000 job titles. The

	occupational classification is structured in a system 3-digit codes and comprises about 330 values. Employment notifications with an end date later than 30 November 2011 are reported using the new occupational classification 2010 (KldB2010). These values are recoded to the KldB1988 by transferring the key area. This results in inaccuracies. **2) ASU, XASU, MTH** The variable documents the occupation of the last job. See 1) for the determination of the occupation.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure. In the XASU source, the occupation variable is not reported for almost the entire period available.
Anonymization	For data privacy reasons, the 3-digit level of the KldB 1988 occupational classification (beruf1988_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the KldB 1988 occupational classification (beruf1988_2).

4.5.4 KldB 2010, Occupation main group – current/most recent (beruf2010_2)

Variable label	KldB 2010, Occupation main group - current/most recent, 2-dig. level
Variable name	beruf2010_2
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	**1) BeH** The occupational title of the job performed by the employee during the notification period is a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes apply for one employee, the employer is required to select the job title that best defines the main activity performed (see Bundesagentur für Arbeit, 2005, p. V). For this, the employer encodes the employee's job in accordance with the "Classification of Occupations 2010" (Klassifikation der Berufe 2010, KldB2010, Bundesagentur für Arbeit, 2011). The occupational class consists of a 5-digit code and comprises about 1,300 values. The occupational group is defined in the first three digits of the code. The skill level required for a job, which is recorded in the fifth digit of the codes in the KldB2010, is made available separately in the variable 'level of requirement' (niveau). Employment notifications with an end date earlier than 30 November 2011 are reported using the occupational classification KldB 1988. These values are recoded to the KldB2010 by transferring the key area. The KldB 2010 is more detailed than the outdated KldB 1988, which causes potential inconsistencies when using the concordance across the classifications. This must be taken into account when analyzing the data.

	2) LeH, ASU, XASU, MTH The variable documents the occupation of the last job. See 1) for the determination of the occupation.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure.
Anonymization	For data privacy reasons, the 3-digit level of the KldB 2010 occupational classification (beruf2010_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the KldB 2010 occupational classification (beruf2010_2).

4.5.5 KldB 2010, Occupation group – current/most recent (beruf2010_3)

Variable label	KldB 2010, Occupation group - current/most recent, 3-digit level
Variable name	beruf2010_3
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH, ASU, XASU, MTH
Data type	numerical
Hierarchy	none
Detailed description	**1) BeH** The occupational title of the job performed by the employee during the notification period is a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes apply for one employee, the employer is required to select the job title that best defines the main activity performed (see Bundesagentur für Arbeit, 2005, p. V). For this, the employer encodes the employee's job in accordance with the "Classification of Occupations 2010" (Klassifikation der Berufe 2010, KldB2010, Bundesagentur für Arbeit, 2011). The occupational class consists of a 5-digit code and comprises about 1,300 values. The occupational group is defined in the first three digits of the code. The skill level required for a job, which is recorded in the fifth digit of the codes in the KldB2010, is made available separately in the variable 'level of requirement' (niveau). Employment notifications with an end date earlier than 30 November 2011 are reported using the occupational classification KldB 1988. These values are recoded to the KldB2010 by transferring the key area. The KldB 2010 is more detailed than the outdated KldB 1988, which causes potential inconsistencies when using the concordance across the classifications. This must be taken into account when analysing the data. **2) LeH, ASU, XASU, MTH** The variable documents the occupation of the last job. See 1) for the determination of the occupation.

Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure.
Anonymization	For data privacy reasons, the 3-digit level of the KldB 2010 occupational classification (beruf2010_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the KldB 2010 occupational classification (beruf2010_2).

4.5.6 KldB 2010, Level of requirement – current/most recent (niveau)

Variable label	Level of requirement - current/most recent
Variable name	niveau
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH, ASU, XASU, MTH
Data type	Numerical
Hierarchy	None
Detailed description	1) **BeH** The occupational title of the job performed by the employee during the notification period is a component of the 'employment details' submitted by the employer. If more than one job title with different classification codes apply for one employee, the employer is required to select the job title that best defines the main activity performed (see Bundesagentur für Arbeit, 2005, p. V). For this the employer encodes the employee's job in accordance with the "Classification of Occupations 2010" (Klassifikation der Berufe 2010, KldB2010, Bundesagentur für Arbeit, 2011). The occupational class consists of a 5-digit code and comprises about 1,300 values. The skill level required for a job, is recorded in the fifth digit of the KldB2010 and documented in this variable (niveau). Employment notifications with an end date earlier than 30 November 2011 are reported using the occupational classification KldB 1988. These values are recoded to the KldB2010 by transferring the key area. The KldB 2010 is more detailed than the outdated KldB 1988, which causes potential inconsistencies when using the concordance across the classifications. This must be taken into account when analysing the data. **2) LeH, ASU, XASU, MTH** The variable documents the skill level as derived from the occupational code of the last job. See 1) for the determination of the occupation.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure.

4.5.7 Part-time (teilzeit)

Variable label	Part-time
Variable name	teilzeit
Category	information on employment, benefit receipt and job search
Origin	BeH

Data type	Numerical
Hierarchy	None
Detailed description	The employee's occupational status during the notification period is reported by the employer in the "employment details". The variable "occupational status" distinguishes between full-time and part-time employees. The determination follows the ratio between the contracted hours and the usual working hours in the establishment. For part-time employees the variable only records whether their working hours exceed a certain limit or not. Until 1978 this limit was 20 hours of work per week, between 1979 and 1987, it was 15 hours per week and since 1988 it has been 18 hours per week. The variable only provides actual details regarding the occupational status for full-time employees, distinguishing among other things between blue-collar and white-collar employees in full-time employment and trainees/apprentices. With the implementation of KldB 2010, however, this distinction is no longer available. The variable 'teilzeit' therefore only distinguishes between full-time and part-time employment in the entire reporting period. No further information about the occupational status is used.
Notes on quality	There is a considerable increase in the number of missing values in 2011 that is due to the change in the reporting procedure. In order to reduce this problem, actual working hours were imputed at the IAB for the period in question. The imputation procedures are described in Ludsteck and Thomsen (2016).

4.5.8 Employment status (erwstat)

Variable label	Employment status
Variable name	erwstat
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	numerical
Hierarchy	none
Detailed description	The content of the variable differs across the data sources. **1) BeH** For BeH observations, the variable 'employment status' corresponds to the so-called person group variable recorded in the notification procedure (DEÜV) that is valid since 1 January 1999. It indicates contribution- or benefit-related particularities of the employment relationship. If multiple codes apply to a single employment notification, the code that prioritized in the official hierarchy must be indicated by the reporting employer. The majority of these cases are jobs subject to social security contributions without any special characteristics (i.e. regular jobs), which are recorded with the value 101. Thus, it is possible that this job type are slightly overestimated.

	The notification procedure stipulates that changes in the employment status - e.g. when an apprentice is retained by his/her training firm after completion of vocational training - must be indicated by a new notification (e.g. Deutsche BKK 2012, p. 31). The so-called person group can be contained in employment notifications that refer to the years prior to 1999 but were not received until 1999 or later. For notifications that were filed prior to 1999 and therefor lack the actual person group code, a concordance is employed to extrapolate and harmonize the information over time. This concordance uses information on vocational education and training, 'occupational status and working hours' and 'occupation' as well as other information that is very specific to categories of the person code classification scheme. Since 1 April 1999, employees in marginal part-time employment have also been recorded in the DEÜV notification procedure. These employees are coded with the values 109 and 209. For employees in marginal part-time employment, no data prior to the introduction of the notification obligation in 1999 could be collected. **2) LeH** For LeH observations, the variable 'employment status' contains the grouped information on the type of benefit these individuals received from the social security system based on the SGB III. Thus, it is possible to distinguish whether a person receives unemployment benefit, unemployment assistance or maintenance allowance or whether contributions to private long-term care insurance are paid by the BA. **3) ASU/XASU** For ASU observations, the 'employment status' variable reports the job search status. Recipients of unemployment benefits (Unemployment Benefit I or II) exceeding the age of 58 years who receive benefits under the relaxed conditions according to Section 428 of Social Code Book III (or Section 65 Para. 4 of Social Code Book II) and individuals aged over 58 years who are not benefit recipients and are not willing to be placed in employment in the sense of Section 252 Paragr. 8 Social Code Book VI are recorded as individuals seeking advice. Individuals recorded as 'without status' (statistics: 'not set') are mainly individuals who cannot be expected to be activated or placed in employment following the definition laid down Section 10 SGB II. Individuals who are classified as unfit for work for more than 42 days but continue to receive Unemployment Benefit II are also classified into the group 'without status'. In XASU observations, the variable 'employment status' has so far also contained the values 'not unemployed, but seeking work' as well as 'unemployed and simultaneously seeking work'. **4) LHG** For LHG datasets, the 'employment status' variable indicates whether the person is registered as an employable minor, an employable person of full age or not employable from the old-age pension threshold.

	5) MTH No detailed information on actual measures is available in the INV-BIO data set. The information recorded on measures in erwstat is aligned with the information that is recorded in the variable quelle (source of data).

4.5.9 Transition zone (gleitz)

Variable label	Transition zone
Variable name	gleitz
Category	information on employment, benefit receipt and job search
Origin	BeH
Data type	numerical
Hierarchy	none
Detailed description	This variable is only available since 2003 and only for BeH observations. It indicates whether the employment notification relates to employment in the low-wage sector, within the so-called transition zone. Jobs in the transition zone have a gross monthly wage of € 400.01 to € 800.00 (so-called midi jobs) for which the employee only has to pay a reduced overall social security contribution. As employees with earnings in the transition zone can voluntarily pay the “regular” social security contribution, not all employees with corresponding earnings are automatically classified as being in the transition zone. The corresponding legislation has been in force since 1 April 2003.

4.5.10 Temporary agency work (leih)

Variable label	Temporary agency work
Variable name	leih
Category	information on employment, benefit receipt and job search
Origin	BeH
Data type	numerical
Hierarchy	none
Detailed description	The variable reports whether the person’s employment is a temporary job via an employment agency. The variable is derived from the occupation code 2010 and is only available for records with an end date later than 30 November 2011.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure.

4.5.11 Fixed-term contract (befrist)

Variable label	Fixed-term contract
Variable name	befrist
Category	information on employment, benefit receipt and job search
Origin	BeH
Data type	numerical

Hierarchy	none
Detailed description	The variable reports whether the person's employment relationship is fixed-term or permanent. The variable is derived from the occupation code 2010 and is only available for notifications with an end date later than 30 November 2011.
Notes on quality	There is a considerable increase in the number of missing values in 2011 due to the change in the reporting procedure.

4.5.12 Reason of cancellation/ notification/ termination (grund)

Variable label	Reason of cancellation/ notification/ termination
Variable name	grund
Category	information on employment, benefit receipt and job search
Origin	BeH, LeH, LHG, ASU, XASU
Data type	numerical
Hierarchy	none
Detailed description	**1) BeH** For BeH records, the 'reason for notification' variable indicates the reason why the employer (establishment) submitted the focal employment notification to the social security agencies. However, not all of the possible reasons for submitting a notification that may occur in the context of the notification procedure are available in the IEB. For instance, the IEB only includes notifications that include valid information on earnings (i.e. annual, employment interruption and end of employment notifications). Initial registrations are not contained as they contain no information on earnings. However, this does not result in a loss of information, as the information is reported again with the subsequent annual, employment interruption or end of employment notification. The reasons for submitting employment notifications are encoded following to the regulations of the official notification procedure, which has been in effect since 1 January 1999 (in accordance with DEÜV). **3) LHG** LHG records contain the 'reason for discontinuation of Unemployment Benefit II' and indicate the reason why current benefits have been discontinued. The 'reason for discontinuation of Unemployment Benefit II' variable refer to the individual and not to the benefit community, which might involve additional individuals. If the Unemployment Benefit II receipt of a different member of the benefit community is discontinued, new observations for all members of the benefit community are started on this date, but the reason for discontinuation of Unemployment Benefit II is only available for the individual whose benefit is discontinued. This variable is valid exactly at the end of the original record. **4) ASU**

	For ASU observations, the variable documents the reason for deregistration. Changes of the legal context in which job search occurs (e.g. an employee who is seeking advice from the BA becomes unemployed) cause the record to be splitted and 'generated by data splitting' is recorded as the reason for deregistration. The number of values of the variable was reduced after 26 April 2003. For analyses over long periods of time, the old values can be recoded to the currently valid ones using the table below: **old -> new old -> new old -> new old -> new** 29 -> 60 36 -> 61 44 -> 74 51 -> 74 30 -> 60 37 -> 66 45 -> 77 52 -> 76 31 -> 61 38 -> 66 46 -> 67 53 -> 68 32 -> 60 39 -> 71 47 -> 67 54 -> 78 33 -> 60 40 -> 69 48 -> 78 34 -> 60 42 -> 65 49 -> 69 35 -> 60 43 -> 70 50 -> 75 **5) XASU** For XASU observations, the variable documents the reason for deregistration. Changes of the legal context in which job search occurs (e.g. an employee who is seeking advice from the BA becomes unemployed) cause the record to be splitted and 'generated by data splitting' is recorded as the reason for deregistration.
Notes on quality	The proportion of valid information (degree of completeness) for the reason for notification in the LHG data sources is very limited (< 30%) in all available years.

4.5.13 Start date of unemployment (alo_beg)

Variable label	Start date of unemployment
Variable name	alo_beg
Category	information on employment, benefit receipt and job search
Origin	LeH, LHG, ASU, XASU, MTH
Data type	numerical
Hierarchy	none
Detailed description	The variable reports the start date of an uninterrupted sequence of periods of unemployment and is calculated at the beginning of the observation. The following gaps do not result in an interruption of the period of unemployment: • any gap lasting seven days or less • periods of illness lasting up to 42 days (ASU) No information about illnesses is contained in XASU observations, which is why it cannot be taken into account in the calculations.

4.5.14 Duration of unemployment (alo_dau)

Variable label	Duration of unemployment

Variable name	alo_dau
Category	information on employment, benefit receipt and job search
Origin	LeH, LHG, ASU, XASU, MTH
Data type	numerical
Hierarchy	none
Detailed description	The variable reports the duration (in days) of an uninterrupted sequence of periods of unemployment and is calculated at the beginning of the observation. The following gaps do not result in an interruption of the period of unemployment: • any gap lasting seven days or less • periods of illness lasting up to 42 days (ASU) When calculating the duration these gaps are not added, however. No information about illnesses is contained in XASU observations, which is why it cannot be taken into account in the calculations. Prior to 1997 the value “0” does not mean that the individual was not unemployed, as the ASU/XASU sources are not available here.

4.6 Location data

4.6.1 Place of residence: district (Kreis/ NUTS 3) (wo_kreis)

Variable label	Place of residence: district (Kreis/ NUTS 3)
Variable name	wo_kreis
Category	location data
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	numerical
Hierarchy	federal state | district
Detailed description	The variable documents the district in which the social security contributor lives in the official AGS classification scheme. The first two digits of the code scheme identify the 16 federal states (Bundesland), digits 1-3 identify regional administrative units in some states (Regierungsbezirk) and digits 1-5 identify the district (Landkreise und kreisfreie Städte). Federal states without regional administrative units have a 0 in the third position. This hierarchy of administrative units is aligned with the NUTS territorial system of EUROSTAT. Landkreise and Kreisfreie Städte correspond to the NUTS 3 hierarchy level as defined by EUROSTAT. BeH records include the place of residence of the employee that is determined at the end of the focal year or at the end of the record if the job is terminated earlier in the year. For records in LHG, XLHG and XASU, the place of residence applies to the period of the original observation. For the ASU and LeH, the variable contains the place of residence at the beginning of the original period of time.

	In order to guarantee consistent regional allocations across the entire observation period, the information on the district was recoded with reference to the territorial allocation of 31 December 2016 for all sources, i.e. in all calendar years, a place of residence is assigned to a district in accordance with the boundaries that the district / NUTS 3 region had on 31 December 2016. As the district boundaries have significantly changed over time due to various reforms of the administrative units, cases would occur in which the district code changes without the individual concerned having relocated if the territorial allocations of the districts were not updated.
Notes on quality	For BeH and LeH records, the place of residence at district level is only available since the year 1999.
Anonymization	For data privacy reasons, the full detail of the classification (wo_kreis) is only available on request and in well-founded cases. By default, the INV-BIO data include wo_bula (16 federal states / NUTS 1 regions).

4.6.2 Place of residence: federal state (Bundesland/ NUTS 1) (wo_bula)

Variable label	Place of residence: federal state (Bundesland/ NUTS1)
Variable name	wo_bula
Category	location data
Origin	BeH, LeH, LHG, ASU, XASU, MTH
Data type	numerical
Hierarchy	federal state | district
Detailed description	This variable is derived from the variable wo_kreis. Its first two digits uniquely identify the 16 federal states/ NUTS 1 regions in Germany. Federal States in Germany correspond to the NUTS 1 hierarchy level as defined by EUROSTAT. For further details, see the description of the variable “Place of residence: district (Kreis/ NUTS 3) (wo_kreis)” (see section 4.6.1).
Anonymization	For data privacy reasons, the full detail of the classification (wo_kreis) is only available on request and in well-founded cases. By default, the INV-BIO data include wo_bula (16 federal states / NUTS 1 regions).

4.7 Establishment characteristics

4.7.1 German classification of economic activity WS 1973, 2-digit level (w73_2)

Variable label	WS 1973 classification, 2-dig. level
Variable name	w73_2
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	division (1-digit code) | group (2-digit code) | class (3-digit code) of economic activity

Detailed description	This variable indicates the economic activity as a 2-digit code following with the WS 1973 classification scheme and is available for the period 1975 up to and including 2002. WS 1973 is the “Classification of Economic Activities for the Statistics of the Federal Employment Services, edition 1973” (“Klassifikation der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit, Ausgabe 1973”). The first digit of the code defines the division of economic activity (N=10), and the first two digits define the 95 groups of economic activity. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.
Anonymization	For data privacy reasons, the 3-digit level of the WS 1973 classification (w73_3) is only available on request and in well-founded cases. By default, the INV-BIO data include w73_2, the 2-digit level of WS 1973 classification.

4.7.2 German classification of economic activity WS 1973, 3-digit level (w73_3)

Variable label	WS 1973 classification, 3-dig. level
Variable name	w73_3
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	division (1-digit code) | group (2-digit code) | class (3-digit code) of economic activity
Detailed description	This variable indicates the economic activity as a 3-digit code following with the WS 1973 classification scheme and is available for the period 1975 up to and including 2002. WS 1973 is the “Classification of Economic Activities for the Statistics of the Federal Employment Services, edition 1973” (“Klassifikation der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit, Ausgabe 1973”). Using a 3-digit code, the classification distinguishes between 269 classes of economic activity. The first digit of the code defines the division of economic activity (N=10), and the first two digits define the 95 groups of economic activity. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.

Anonymization	For data privacy reasons, the 3-digit level of the WS 1973 classification (w73_3) is only available on request and in well-founded cases. By default, the INV-BIO data include w73_2, the 2-digit level of WS 1973 classification.

4.7.3 NACE Rev. 1 / German classification of economic activity WZ 1993, 2-digit level (w93_2)

Variable label	NACE Rev. 1 / WZ 1993 classification, 2-dig. level
Variable name	w93_2
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable indicates the economic activity as a 2-digit code following the WZ 1993 classification scheme and is available from 1999 up to and including 2003. WZ 1993 is the “Classification of Economic Activities for the Statistics of the Federal Employment Services, edition 1993” (“Klassifikation der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit, Ausgabe 1993”). The WZ 1993 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 1 („Nomenclature génerale des activités économiques dans les communautés européennes“). Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_w93_2).

4.7.4 NACE Rev. 1 / German classification of economic activity WZ 1993, 3-digit level (w93_3)

Variable label	NACE Rev. 1 / WZ 1993 classification, 3-dig. level
Variable name	w93_3
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity

Detailed description	This variable indicates the economic activity as a 3-digit code following with the WZ93 classification scheme and is available from 1999 up to and including 2003. WZ 1993 is the “Classification of Economic Activities for the Statistics of the Federal Employment Services, edition 1993” (“Klassifikation der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit, Ausgabe 1993”). The WZ 1993 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 1 („Nomenclature génerale des activités économiques dans les communautés européennes“). Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_w93_2).

4.7.5 NACE Rev. 1.1 / German classification of economic activity WZ 2003, 2-digit level (w03_2)

Variable label	NACE Rev. 1.1 / WZ 2003 classification, 2-dig. level
Variable name	w03_2
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5-digit code) of economic activity
Detailed description	This variable indicates the economic activity as a 2-digit code following with the WZ 2003 classification scheme and is available from 2003 up to 2008. WZ 2003 is the “Classification of Economic Activities, Edition 2003” (“Klassifikation der Wirtschaftszweige Ausgabe 2003”) of the Federal Statistical Office (eds.). The WZ 2003 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 1.1, which is a minor revision of the NACE Rev. 1 scheme. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.

Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1.1/ German classification of economic activity WZ 2003 (w03_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1.1/ German classification of economic activity WZ 2003 (w03_2).

4.7.6 NACE Rev. 1.1 / German classification of economic activity WZ 2003, 3-digit level (w03_3)

Variable label	NACE Rev. 1.1 / WZ 2003 classification, 3-dig. level
Variable name	w03_3
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable indicates the economic activity as a 3-digit code following with the WZ 2003 classification scheme and is available from 2003 up to 2008. WZ 2003 is the "Classification of Economic Activities, Edition 2003" ("Klassifikation der Wirtschaftszweige Ausgabe 2003") of the Federal Statistical Office (eds.). The WZ 2003 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 1.1, which is a minor revision of the NACE Rev. 1 scheme. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1.1/ German classification of economic activity WZ 2003 (w03_3) is only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1.1/ German classification of economic activity WZ 2003 (w03_2).

4.7.7 NACE Rev. 2 / German classification of economic activity WZ 2008, 2-digit level (w08_2)

Variable label	NACE Rev. 2 / WZ 2008 classification, 2-dig. level
Variable name	w08_2
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5-digit code) of economic activity

Detailed description	This variable indicates the economic activity as a 2-digit code in following the WZ 2008 classification scheme and is available since 2008. WZ08 is for the “Classification of Economic Activities, Edition 2008” (“Klassifikation der Wirtschaftszweige Ausgabe 2008”) of the Federal Statistical Office (eds.). The WZ 2008 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 2, which is a major revision of the NACE Rev. 1.1 classification scheme. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.8 NACE Rev. 2 / German classification of economic activity WZ 2008, 3-digit level (w08_3)

Variable label	NACE Rev. 2 / WZ 2008 classification, 3-dig. level
Variable name	w08_3
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable records the economic activity as a 3-digit code in following the WZ 2008 classification scheme and is available since 2008. WZ 2008 is for the “Classification of Economic Activities, Edition 2008” (“Klassifikation der Wirtschaftszweige Ausgabe 2008”) of the Federal Statistical Office (eds.). The WZ08 scheme is based on the Statistical Classification of Economic Activities in the European Community NACE Rev. 2, which is a major revision of the NACE Rev. 1.1 classification scheme. Each establishment is only assigned one code. The assignment to the relevant class of economic activity is determined from the economic activity performed by the majority of the employees in the unit. Note that economic activity determined at the level of establishments with this reporting directive of the BA does not necessarily conform to the activity that is determined from e.g. sales information in firm level data.

Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.9 NACE Rev. 1 / German classification of economic activity WZ 1993, time consistent, 2-digit level (w93_2_gen)

Variable label	NACE Rev. 1 / WZ 1993 classification, time consistent, 2-dig. level
Variable name	w93_2_gen
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable records the economic activity as a 2-digit code following the WZ 1993 classification scheme. For further information on the WZ 1993 classification scheme, see the description of variable w93_2. For the period 1998-2003, the variable contains the original values recorded in w93_2. Prior to 1998 and after 2003, the original information is either extrapolated (forward/backward) for the same establishment or imputed using concordance tables that document the sectoral transitions of establishments by industry. For further information on the methodology and the concordance tables, see Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_ w93_2).

4.7.10 NACE Rev. 1 / German classification of economic activity WZ 1993, time consistent, type of imputation (group_w93_2)

Variable label	Type of imputation w93_2
Variable name	group_w93_2
Category	establishment variables
Origin	BeH
Data type	Numerical
Hierarchy	None
Detailed description	This variable indicates the type of completion for the w93_2_gen variable. It documents whether the value of w93_2_gen is consistent with the original value from w93_2, still missing/ extrapolated or imputed based on recording tables.

	A detailed description of the procedure can be found in Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_ w93_2).

4.7.11 NACE Rev. 1 / German classification of economic activity WZ 1993, time consistent, 3-digit level (w93_3_gen)

Variable label	NACE Rev. 1 / WZ 1993 classification, time consistent, 3-dig. level
Variable name	w93_3_gen
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable records the economic activity as a 3-digit code following the WZ 1993 classification scheme. For further information on the WZ 1993 classification scheme, see the description of variable w93_3. For the period 1998-2003, the variable contains the original values recorded in w93_3. Prior to 1998 and after 2003, the original information is either extrapolated (forward/backward) for the same establishment or imputed using concordance tables that document the sectoral transitions of establishments by industry. For further information on the methodology and the concordance tables, see Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_ w93_2).

4.7.12 NACE Rev. 1 / German classification of economic activity WZ1993, time consistent, type of imputation (group_w93_3)

Variable label	Type of imputation w93_3
Variable name	group_w93_3
Category	establishment variables
Origin	BeH
Data type	Numerical
Hierarchy	None

Detailed description	This variable indicates the type of completion for the w93_3_gen variable. It documents whether the value of w93_3_gen is consistent with the original value from w93_3, still missing/ extrapolated or imputed based on recording tables. A detailed description of the procedure can be found in Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_3) as well as the time consistent version of this variable (w93_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 1/ German classification of economic activity WZ 1993 (w93_2, w93_2_gen, group_ w93_2).

4.7.13 NACE Rev. 2 / German classification of economic activity WZ 2008, time consistent, 2-dig. level (w08_2_gen)

Variable label	NACE Rev. 2 / WZ 2008 classification, time consistent, 2-dig. level
Variable name	W08_2_gen
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable records the economic activity as a 2-digit code in accordance with the WZ 2008 classification. Since 2008, the variable contains the original values from w08_2. Prior to 2008, the original information is either extrapolated (backward) for the same establishment or imputed using concordance tables that document the sectoral transitions of establishments by industry. For further information on the methodology and the concordance tables see Eberle et al. (2011). Further information on the WZ 2008 classification can be found in the description of variable w08_2.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.14 NACE Rev. 2 / German classification of economic activity WZ 2008, time consistent, type of imputation (group_w08_2)

Variable label	Type of imputation w08_2
Variable name	group_w08_2
Category	establishment variables
Origin	BeH
Data type	numerical

Hierarchy	None
Detailed description	This variable indicates the type of completion for the w08_2_gen variable. It documents whether the value of w08_2_gen is consistent with the original value from w08_2, still missing/ extrapolated or imputed based on recording tables. A detailed description of the procedure can be found in Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.15 NACE Rev. 2 / German classification of economic activity WZ 2008, time consistent, 3-digit level (w08_3_gen)

Variable label	NACE Rev. 2 / WZ 2008 classification, time consistent, 3-dig. level
Variable name	W08_3_gen
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	section (1-digit code) | division (2-digit code) | group (3-digit code) | class (4-digit code) | sub-class (5 digit code) of economic activity
Detailed description	This variable records the economic activity as a 3-digit code in accordance with the WZ 2008 classification. Since 2008, the variable contains the original values from w08_3. Prior to 2008, the original information is either extrapolated (backward) for the same establishment or imputed using concordance tables that document the sectoral transitions of establishments by industry. For further information on the methodology and the concordance tables see Eberle et al. (2011). Further information on the WZ 2008 classification can be found in the description of variable w08_3.
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.16 NACE Rev. 2 / German classification of economic activity WZ 2008, time consistent, type of imputation (group_w08_3)

Variable label	Type of imputation w08_3
Variable name	group_w08_3
Category	establishment variables
Origin	BeH

Data type	numerical
Hierarchy	None
Detailed description	This variable indicates the type of completion for the w08_3_gen variable. It documents whether the value of w08_3_gen is consistent with the original value from w08_3, still missing/ extrapolated or imputed based on recording tables. A detailed description of the procedure can be found in Eberle et al. (2011).
Anonymization	For data privacy reasons, the 3-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_3) as well as the time consistent version of this variable (w08_3_gen) are only available on request and in well-founded cases. By default, the INV-BIO data include the 2-digit level of the NACE Rev. 2/ German classification of economic activity WZ 2008 (w08_2, w08_2_gen, group_ w08_2).

4.7.17 Year of first appearance of establishment number (grd_jahr)

Variable label	year of first appearance
Variable name	grd_jahr
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	none
Detailed description	This variable indicates the first appearance of the establishment number in the dataset. If an establishment number in western Germany is only determined for the first time after 1975 (or after 1992 in eastern Germany), this variable could indicate the date when the respective establishment was founded. However, it could also be an establishment that has been in existence for a longer time but has been allocated a new establishment number following a change of owner or a change in the legal form of the establishment. (For the allocation of establishment numbers see Bundesagentur für Arbeit 2007, pp. 9-11). It is also possible that the establishment already existed before, but had no employees subject to social security, or from 1999 onwards, no marginal part-time workers.

4.7.18 Year of last appearance of establishment number (lzt_jahr)

Variable label	year of last appearance
Variable name	lzt_jahr
Category	establishment variables
Origin	BeH
Data type	numerical
Hierarchy	none

Detailed description	This variable indicates the last appearance of the establishment number in the dataset (see Bender et. al. 1996). If the existence of an establishment number in the BHP already ends before 2008, it could indicate the closure of the establishment. However, other possible reasons for this are an “arbitrary change of the establishment number following a change of owner or a change in the legal form of the establishment”, the “outsourcing of parts of the firm under a new establishment number” or other administrative changes (see Bender et. al. 1996 or Bundesagentur für Arbeit 2007, pp. 9-11).

4.7.19 Total number of employees (az_ges)

Variable label	no. employees
Variable name	az_ges
Category	generated establishment variables
Origin	BeH
Detailed description	This variable contains the total number of an establishment’s employees reported to the social security agencies as of 30 June of a year. Since the introduction of the new notification regulations in 1999, people in marginal part-time employment have also been recorded. Dormant employment relationships (daily wage of zero) are not included. For further details see Schmucker et al. (2016).

4.7.20 Number of full-time employees (regular workers + others) (az_vz)

Variable label	No. full-time (regular workers + others)
Variable name	az_vz
Category	generated establishment variables
Origin	BeH
Hierarchy	none
Detailed description	The variable contains the number of individuals in the establishment who are reported with the person group codes 101, 140, 143, 105, 106, 112, 118, 119, 120, 149, 201, 203, 205, 999 and YYY and as full-time employees. This means that trainees/apprentices and people in marginal part-time employment or in partial retirement are not taken into account. For further details see Schmucker et al. (2016).

4.7.21 Number of employees in marginal part-time employment (az_gf)

Variable label	no. marginal part-time workers
Variable name	az_gf
Category	generated establishment variables
Origin	BeH
Hierarchy	none
Detailed description	The number of marginal part-time employees is generated from the person group code – values 109 and 209. This variable only contains valid values in the dataset since 1999 as people in marginal part-time employment were only integrated into the social security notification procedure from that year onwards. For further details see Schmucker et al. (2016).

4.7.22 Mean imputed wage all full-time employees (te_imp_mw)

Variable label	Mean imp. wage all full-time employees
Variable name	te_imp_mw
Category	generated establishment variables
Origin	BeH
Data type	numerical
Hierarchy	none
Detailed description	This variable contains the mean imputed gross daily wage of the full-time employees in an establishment. It does not include the wages of marginally part-time staff, apprentices or individuals participating in partial retirement schemes. The values are reported in euros for all years. According to the social security notification regulations, employers must indicate the employee's gross wage subject to social security contributions for a certain period of time (fixed period wage). Until the end of 1998, employers had to report the gross wage subject to social security contributions only. So only wages above the marginal part-time income threshold and below the contribution assessment ceiling were recorded. Since 1999, wages below the marginal part-time income threshold have also been recorded as part of the new notification procedure. Gross wages above the contribution assessment ceiling, however, are still cut. In order to calculate the gross daily wage, the fixed period wage is divided by the number of calendar days in the period. To calculate the mean, these censored wages were imputed (see Section 8.2 in Schmucker et al. 2016). These data were then aggregated at establishment level. For further details see Schmucker et al. (2016).

4.7.23 Place of work: district (Kreis/ NUTS 3) (ao_kreis)

Variable label	Place of work: district (Kreis/ NUTS 3)
Variable name	ao_kreis
Category	location data
Origin	BeH
Data type	numerical
Hierarchy	federal state | district
Detailed description	The variable indicates the (administrative) district (Landkreise and Kreisfreie Städte) in which the establishment of the employee is located. The first two digits of the 5-digit code indicate the 16 unique federal states (Bundesland), digitis 1-3 indicate regional administrative units in some states (Regierungsbezirk), and digits 1-5 uniquely identify the district (Kreis) as administrative unit. Federal states without a regional administrative units have a 0 in the third position. Landkreise and

	Kreisfreie Städte correspond to the NUTS 3 hierarchy level as defined by EUROSTAT. In order to guarantee consistent regional allocations across the entire observation period, the information on the district/NUTS 3 region was recoded to the territorial allocation of 31 December 2016, i.e. in all calendar years, a place of work is assigned to a district in accordance with the boundaries that the district/NUTS 3 had on 31 December 2016. As the district boundaries have changed over time, cases would occur in which the district code of the location of the establishment would change without the establishment concerned having relocated, if the territorial allocations were not updated.
Anonymization	For data privacy reasons, the full detail of the classification (ao_kreis) is only available on request and in well-founded cases. By default, the INV-BIO data include ao_bula (16 federal states / NUTS 1 regions).

4.7.24 Place of work: federal state (Bundesland/ NUTS 1) (ao_bula)

Variable label	Place of work: federal state (Bundesland / NUTS 1)
Variable name	ao_bula
Category	location data
Origin	BeH
Data type	numerical
Hierarchy	federal state | district
Detailed description	The variable indicates the federal state in which the establishment is located. This variable is derived from the district code (ao_kreis). The first two digits of the district code uniquely identify the frim location at the level of federal states. Federal states in Germany correspond to the NUTS 1 hierarchy level as defined by EUROSTAT. For further details, see the description of the variable “Place of work: district (Kreis/ NUTS 3) (ao_kreis)” (see section 4.7.23).
Anonymization	For data privacy reasons, the full detail of the classification (ao_kreis) is only available on request and in well-founded cases. By default, the INV-BIO data include ao_bula (16 federal states / NUTS 1 regions).

4.8 Patent characteristics

4.8.1 DOCDB family ID (docdb_family_id)

Variable label	DOCDB family ID
Variable name	family_id
Category	Identifiers
Origin	PATSTAT (Version 10/2017)
Data type	Numerical
Hierarchy	None
Detailed description	Anonymized unique DOCDB patent family ID. The variable is a pseudonymized equivalent to docdb_family_id that is included in the PATSTAT data.

	Patent families represent a group of patent applications protecting the same technology as defined by the patent examiners (see e.g. Martinez 2010; EPO 2018). Thus, patent families are considered to cover a single invention. Members of a DOCDB patent family all have exactly the same priorities.
Notes on quality	Within each inventor biography, a patent family is represented by the earliest patent application filed by the inventor. Please note that within patent families (i.e. multiple patent filings) the set of inventors may vary in rare cases. Therefore, also the earliest application of the inventor within the family that represents the DOCDB family may vary across inventors.

4.8.2 Patent application ID, earliest matched appl. within DOCDB family by inventor (appln_id)

Variable label	Patent application ID, earliest matched appl. within DOCDB family by inventor
Variable name	appln_id
Category	identifiers
Origin	Patent register data
Data type	numerical
Hierarchy	DOCDB family ID
Detailed description	Anonymized unique patent applicant ID. The variable is a pseudonymized equivalent to appln_id that is included in the PATSTAT data. Appln_id identifies unique patent applications, which might be nested in DOCDB patent families (see section 3.1).

4.8.3 Number of applications within DOCDB family (docdb_family_size)

Variable label	Number of applications within DOCDB family
Variable name	docdb_family_size
Category	identifiers
Origin	Patent register data
Data type	numerical
Hierarchy	DOCDB family ID
Detailed description	The variable documents the number of patent applications (appln_id) that belong to the DOCDB patent family. Note that not all of these patent applications (appln_id) are recorded in the INV-BIO data because patent families may include patent applications filed with different patent offices or jurisdictions, including offices that are by definition excluded from our sample (e.g. US, JP, etc.). Since international patent filings are relatively expensive, the literature usually interprets patent family size as a correlate of patent value (e.g. Lanjouw et al. 1998; Harhoff et al. 2003).

4.8.4 Earliest application filing date within DOCDB family (earliest_filing_date)

Variable label	Earliest application filing date within DOCDB family
Variable name	earliest_filing_date

Category	Patent characteristics
Origin	Patent register data
Data type	Date (YYYY/Quarter)
Hierarchy	DocDB family ID
Detailed description	Earliest patent application filing date (YYYY/quarter) within DOCDB family. This date is unique across all applications that belong to the same DOCDB family.

4.8.5 Application filing date of earliest patent appl. within DOCDB family by inventor (appln_filing_date)

Variable label	Application filing date of earliest patent appl. within DOCDB family by inventor
Variable name	appln_filing_date
Category	Patent characteristics
Origin	Patent register data
Data type	Date (YYYY/Quarter)
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Earliest matched patent application date within DOCDB family by inventor.

4.8.6 Earliest patent publication date within DOCDB family (earliest_publn_date)

Variable label	Earliest patent publication date within DOCDB family
Variable name	earliest_publn_date
Category	Patent characteristics
Origin	Patent register data
Data type	Date (YYYY/Quarter)
Hierarchy	DOCDB patent family
Detailed description	Earliest patent publication filing date (YYYY/quarter) within DOCDB family. By definition, this date is unique across all applications that belong to same DOCDB family.

4.8.7 Patent application is granted (granted)

Variable label	Granted patent application
Variable name	granted
Category	Patent characteristics
Origin	Patent register data
Data type	numerical (binary)
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Indicates whether the earliest matched application (appln_id) of the focal inventor within a DOCDB patent family was granted by the patent office or not.

	Note that at the application level within DOCDB families, not all patents must be granted. Thus, while being accurate information for the earliest patent application, the variable is less reliable for the DOCDB family as a whole.

4.8.8 Grant date of patent application (grant_date)

Variable label	Grant date of patent application
Variable name	grant_date
Category	Patent characteristics
Origin	Patent register data
Data type	Date (YYYY/quarter)
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Grant date (YYYY/quarter) of earliest matched patent application of the focal inventor within DOCDB family ID. Date is coded systematically missing for patent applications that are not granted.

4.8.9 DOCDB citations within {X} yrs – DE (cit_docdb_DE_{X}yrs)

Variable label	DOCDB forward citations within {X} yrs for jurisdiction DE
Variable name	cit_docdb_DE_{X}yrs
Category	Patent characteristics
Origin	Patent register data
Data type	Numerical
Hierarchy	DOCDB patent family
Detailed description	Number of forward citations that the invention (DOCDB family) received from patent applications at the German Patent and Trademark Office after 2, 3, 4, 5, 6, 7, 8, 9, 10 years from the earliest publication date, respectively. The citation counts are corrected for equivalents across patent authorities and include citations by the applicant. Count variables for each year are truncated at the 99th percentile of the distribution of citations. Citations are assumed to mirror the technological importance for subsequent developments (e.g. Trajtenberg 1990).

4.8.10 DOCDB citations within {X} yrs – EP (cit_docdb_EP_{X}yrs)

Variable label	DOCDB forward citations within {X} yrs for jurisdiction EP
Variable name	cit_docdb_EP_{X}yrs
Category	Patent characteristics
Origin	Patent register data
Data type	Numerical
Hierarchy	DOCDB patent family

Detailed description	Number of forward citations that the invention (DOCDB family) received from patent applications at the European Patent Office after 2, 3, 4, 5, 6, 7, 8, 9, 10 years from the earliest publication date, respectively. The citation counts are corrected for equivalents across patent authorities and include citations by the applicant. Count variables for each year are truncated at the 99th percentile of the distribution of citations. Citations are assumed to mirror the technological importance for subsequent developments (e.g. Trajtenberg 1990).

4.8.11 DOCDB citations within {X} yrs – US (cit_docdb_US_{X}yrs)

Variable label	DOCDB forward citations within {X} yrs for jurisdiction EP
Variable name	cit_docdb_US_{X}yrs
Category	Patent characteristics
Origin	Patent register data
Data type	Numerical
Hierarchy	DOCDB patent family
Detailed description	Number of forward citations that the invention (DOCDB family) received from patent applications at the US Patent and Trademark Office after 2, 3, 4, 5, 6, 7, 8, 9, 10 years from the earliest publication date, respectively. The citation counts are corrected for equivalents across patent authorities and include citations by the applicant. Count variables for each year are truncated at the 99th percentile of the distribution of citations. Citations are assumed to mirror the technological importance for subsequent developments (e.g. Trajtenberg 1990).

4.8.12 Generality measure (generality_docdb)

Variable label	Generality measure based on forward citations to DOCDB family
Variable name	generality_docdb
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	DOCDB patent family
Detailed description	Mirrors number and distribution of forward citations and IPC classes cites belong to. The definition follows Trajtenberg et. al (1997): $\text{Generality}_p = 1 - \sum_j {S_{pj}}^2$, where S_{pj} represents the share of forward cites to patent *p* from class *j* out of 34 IPC tech areas. A higher generality value for a patent tends to be associated with a more valuable invention.

4.8.13 Originality measure (originality_docdb)

Variable label	Originality measure based on backward citations from DOCDB family
Variable name	originality_docdb
Category	Patent characteristics

Origin	Patent register data
Data type	numerical
Hierarchy	DOCDB patent family
Detailed description	Mirrors number and distribution of backward citations and IPC classes cites belong to. The definition follows Trajtenberg et. al (1997): $\text{Originality}_p = 1 - \sum_j {S_{pj}}^2$, where S_{pj} represents the share of backward citations of patent *p* from class *j* out of 34 IPC tech areas. A higher originality value for a patent tends to be associated with a more radical invention.

4.8.14 Number of inventors (nb_inventors)

Variable label	Number of inventors
Variable name	nb_inventors
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Number of unconsolidated inventors listed on earliest matched patent application of the focal inventor within DOCDB family. Note that inventor information may deviate across applications within the same DOCDB family. This is because the number of inventors (as well as their residential address) may change over the course of time. However, variation of inventor information within DOCDB families is generally very low and changes are rare events.

4.8.15 Foreign inventors (d_foreign_inv)

Variable label	Foreign inventors
Variable name	d_foreign_inv
Category	Patent characteristics
Origin	Patent register data
Data type	numerical (binary)
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Indicator variable for patent application whether one or more inventors are listed under foreign (non-“DE”) residential address in the patent register. Note that inventor information may deviate across applications within the same DOCDB family. This is because the number of inventors (as well as their residential address) may change over the course of time. However, variation of inventor information within DOCDB families is generally very low and changes are rare events.

4.8.16 Complete inventor team (pat_complete)

Variable label	Complete inventor team identified as employees at application filing date
Variable name	pat_complete
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Indicator variable for each patent application, documenting whether the full number of inventors listed on the patent (nb_inventors) are matched with employees in the administrative labor market data and at the time of the earliest patent filing, also an employment episode is recorded.

4.8.17 Number of applicants (nb_applicants)

Variable label	Number of applicants
Variable name	nb_applicants
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Number of unconsolidated applicants/ assignees listed on earliest matched patent application of the focal inventor within DOCDB family. Note that the application information may not be representative for other applications within DOCDB family, since the number of applicants (as well as their address) may change over the course of time. Applicant changes (also within DOCDB families) are more frequent than inventor changes and occur if patents are transferred to another applicant (for a survey on the literature on patent transfers, see Gaessler 2016)

4.8.18 Foreign applicants (d_foreign_appl)

Variable label	Foreign applicants
Variable name	d_foreign_appl
Category	Patent characteristics
Origin	Patent register data
Data type	numerical (binary)
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Indicator variable for patent application whether one or more applicants are listed under foreign (non-“DE”) address in patent register.

	Note that the application information may deviate across applications within the same DOCDB family. This is because the number of applicants (as well as their address) may change over the course of time.

4.8.19 Technology area (area34)

Variable label	Technology area
Variable name	area34
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	The technology area classifies the technology of the invention at the patent level into 34 classes. The classification used in the INV-BIO data is a slightly modified version of the WIPO classification (Schmoch 2008) that originally proposes 35 technology areas, however, the classes 21 (Surface technology) and 22 (Nanotechnology) are merged. Technology areas for patents as recorded in the INV-BIO are generated in two steps: 1. The WIPO technology concordance table (Schmoch 2008) is used to merge the technology area code to the detailed technology codes in the IPC classification that are assigned by the patent examiners. 2. Using the full set of transformed area codes, the modal technology area for each patent application is identified. If the modal technology area is ambiguous, a random choice among the candidates with the highest count was made.
Notes on quality	In less than two percent of the DOCB families recorded in the INV-BIO data, the DOCDB modal technology class differs from the technology area recorded at the application level information.

4.8.20 Technology main area (mainarea34)

Variable label	Technology main area
Variable name	mainarea34
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	The technology main area classifies the technology of the invention into 5 aggregate technology classes. Main areas represent groups of the 34 technology areas.

	Technology areas in the INV-BIO data refer to the application level and, thus, to the earliest matched patent application within DOCDB patent family by inventor.
Notes on quality	For only less than two percent of the DOCB families recorded in the data, the DOCDB modal technology class differs from the technology area recorded for the application level information.

4.8.21 Matched inventors are employed with multiple establishments (multi_betnr)

Variable label	Matched inventors are employed with multiple establishments (0/1)
Variable name	multi_betnr
Category	Patent characteristics
Origin	Patent register data
Data type	Numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Indicator variable that takes the value of one if the matched inventors work in different establishments at the earliest application date within the DOCDB family.

4.8.22 Average distance between matched inventors (mean_dist_inv)

Variable label	Average distance between matched inventors (km)
Variable name	mean_dist_inv
Category	Patent characteristics
Origin	Patent register data
Data type	numerical
Hierarchy	Patent application level (appln_id), i.e. earliest matched application within DOCDB patent family by inventor.
Detailed description	Average spatial distance (beeline) between the work locations of the inventors. The distance calculation is based on the centroids of the municipalities in which the establishments of the inventors are located.
Notes on quality	The variable is systematically missing for patents that list only one inventor and/or only one inventor is matched with the administrative labor market data. Also, missing data on the work location affect the quality of the variable. If all inventors are employed with the same establishment or at establishments located in the same municipality, then, by definition, the distance takes the value of zero.

5. References

Aghion, P. / Akcigit, U. / Hyytinen, A. / Toivanen, O. (2017): The Social Origins of Inventors. CEP Discussion Paper 1522. London.

Akcigit, U. / Grigsby, J. / Nicholas, T. (2017): The Rise of American Ingenuity: Innovation and Inventors of the Golden Age. NBER Discussion Paper 23047. Cambridge/MA.

Antoni, M. / Ganzer, A. / vom Berge, P. (2016): Sample of Integrated Labour Market Biographies (SIAB) 1975-2014. FDZ-Datenreport, 04/2016 (en), Nuremberg.

Bachteler, T. (2008): Dokumentation Record Linkage IEB-PASS. Mimeo.

Bell, A. / Chetty, R. / Jaravel, X. / Petkova, N. / Van Reenen, J. (2017): Who Becomes an Inventor in America? The Importance of Exposure to Innovation. NBER Discussion Paper 24062. Cambridge/MA.

Bender, S. / Hilzendegen, J. / Rohwer, G. / Rudolph, H. (1996): Die IAB-Beschäftigtenstichprobe 1975-1990. Beiträge zur Arbeitsmarkt- und Berufsforschung, 197, Nürnberg.

Bertat, T. / Dundler, A. / Grimm, C. / Kiewitt, J. / Schomaker, C. / Schridde, Dr. H. / Zemann, Dr. C. (2013): Neue Erhebungsinhalte ‚Arbeitszeit', ‚ausgeübte Tätigkeit' sowie ‚Schul- und Berufsabschluss' in der Beschäftigungsstatistik. Methodenbericht, Bundesagentur für Arbeit – Statistik, URL: http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Methodenberichte/Beschaeftigungsstatistik/Generische-Publikationen/Methodenbericht-Neue-Erbebungsinhalte-Arbeitszeita-ausgeuebte-Taetigkeit-sowie-Schul-und-Berufsabschluss-in-der-Beschaeftigungsstatistik.pdf, (31 March 2016).

Bundesagentur für Arbeit (2009): Klassifikation der Wirtschaftszweige 1973, Nürnberg. URL: http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Wirtschaftszweige/Klassifikation-der-Wirtschaftszweige-1973-2003/Klassifikationen-der-Wirtschaftszweige-1973-2003-Nav.html, (21 April 2016).

Bundesagentur für Arbeit (2011): Klassifikation der Berufe 2010. Band 1: Systematischer und alphabetischer Teil mit Erläuterungen, Nürnberg. URL: http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/Printausgabe-KldB2010/Printausgabe-KldB-2010-Nav.html, (21 April 2016).

Bundesagentur für Arbeit (Hrsg.) (2005): Schlüsselverzeichnis für die Angaben zur Tätigkeit in den Meldungen zur Sozialversicherung. Ausgabe Januar 2005, Nürnberg.

Bundesagentur für Arbeit (Hrsg.) (2007): Handbuch für die Betriebsnummernvergabe und –pflege im Rahmen des Meldeverfahren zur Sozialversicherung. Ausgabe Dezember 2007, Nürnberg.

Bundesanstalt für Arbeit (1988): Klassifikation der Berufe – Systematisches und Alphabetisches Verzeichnis der Berufsbenennung, Nürnberg. URL:

http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB1975-1992/KldB1975-1992-Nav.html, (21 April 2016).

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

Cramer, U. (1985): Probleme der Genauigkeit der Beschäftigtenstatistik. Allgemeines Statistisches Archiv, 69, 56-68.

Depalo, D. / Di Addario S. (2014): Shedding Light on Inventors' Returns to Patents. IRLE Working Paper No. 115-14, Berkeley/CA.

Deutsche BKK (2016): Ratgeber Sozialversicherung 2016, Wolfsburg, URL: https://www.deutschebkk.de/fileadmin/user_upload/microsites/arbeitgeber/medien/pdf/ratgeber-sozialversicherung-2016.pdf, (31 March 2016).

Dorner, M. / Harhoff, D. (2018): A Novel Technology-Industry Concordance Table based on Linked Inventor Establishment Data. Research Policy, 47 (4), 768-781.

Drews, N. (2006): Qualitätsverbesserung der Bildungsvariable in der IAB-Beschäftigtenstichprobe 1975-2001. FDZ Methodenreport, 05/2006 (de), Nürnberg.

Eberle, J. / Jacobebbinghaus, P. / Ludsteck, J. / Witter, J. (2011): Generation of time-consistent industry codes in the face of classification changes * Simple heuristic based on the Establishment History Panel (BHP). FDZ Methodenreport, 05/2011 (en), Nürnberg.

European Patent Office / EPO (2017): Patent Families at the EPO. Mimeo. URL: http://documents.epo.org/projects/babylon/eponet.nsf/0/C9387E5053AA707BC125816A00508E8D/$File/Patent_Families_at_the_EPO_en.pdf, (26 June 2018).

Fellegi I. / Sunter A. (1969): A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183-1210.

Fitzenberger, B. / Osikominu, A. / Völter, R. (2006): Imputation rules to improve the education variable in the IAB employment subsample. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 126 (3), 405-436.

Gaessler, Fabian (2016): Enforcing and Trading Patents – Evidence for Europe (Innovation und Entrepreneurship). Wiesbaden: Springer Gabler.

Gambardella, A. / Giuri, P. / Mariani, M. (2005): The Value of European Patents: Evidence from a survey of European Inventors – Final report of the PatVal EU Project. http://www.alfonsogambardella.it/PATVALFinalReport.pdf, (12 December 2013).

Ge, C. / Huang, K. / Png, I. P.L. (2016): Engineer/Scientist Careers: Patents, Online Profiles, and Misclassification, Strategic Management Journal, 37 (1), 232-253.

Harhoff, D. / Scherer, F. M. / Vopel, K. (2003): Citations, family size, opposition and the value of patent rights. Research Policy, 32 (8), 1343-1363.

Hastie, T. / Tibshirani, R. / Friedman, J. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), Springer-Verlag. URL: https://web.stanford.edu/~hastie/Papers/ESLII.pdf, (18 June 2018).

Jung, T. / Ejermo, O. (2014): Demographic patterns and trends in patenting: Gender, age, and education of inventors. Technological Forecasting and Social Change, 86, 110-124.

Lanjouw, J. O. / Pakes, A. / Putnam, J. (1998): How to count patents and value intellectual property: The uses of patent renewal and application data. The Journal of Industrial Economics, 46 (4), 405-432.

Li, G. C. / Lai, R. / D'Amour, A. / Doolin, D.M. / Sun, Y. / Torvik, V.I. / Yu, A.Z. / Fleming, L. (2014): Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). Research Policy, 43 (6), 941-955.

Ludsteck, J. / Thomsen, U. (2016): Imputation of the Working Time Information for the Employment Register Data. FDZ Methodenreport 01/2016 (en), Nürnberg.

Martinez, C. (2010): Insight into Different Types of Patent Families. OECD STI Working Paper 2010/2, OECD, Paris.

Meinken, H. / Koch, I. (2004): BA-Beschäftigtenpanel 1998-2002. Codebuch, Nürnberg.

Morrison, G. / Riccaboni, M. / Pammolli, F. (2017): Disambiguation of patent inventors and assignees using high-resolution geolocation data. Nature Scientific data, 4, 170064. URL: https://www.nature.com/articles/sdata201764, (18 June 2018).

Paulus, W. / Matthes, B. (2013): Klassifikation der Berufe * Struktur, Codierung und Umsteigeschlüssel. FDZ-Methodenreport, 08/2013 (de), Nürnberg.

Pezzoni, M. / Lissoni, F. / Tarasconi, G. (2014): How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation. Scientometrics, 101 (1), 477-504.

Raffo, J. / Lhuillery, S. (2009): How to play the "Names Game": Patent retrieval comparing different heuristics. Research Policy, 38 (10), 1617-1627.

Schmoch, U. (2008): Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Organisation (WIPO). Fraunhofer Institute for Systems and Innovation Research, Karlsruhe, Germany. URL: http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf, (18 June 2018).

Schmucker, A.; Seth, S.; Ludsteck, J.; Eberle, J.; Ganzer, A. (2016): Establishment History Panel 1975-2014. FDZ-Datenreport, 03/2016 (en), Nürnberg.

Schnell, R. / Bachteler, T. / Bender, S. (2004): A Toolbox for record linkage; Austrian Journal of Statistics, 33 (1-2), 125-133.

Statistisches Bundesamt (2002): Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ 93), Wiesbaden. URL: https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaftklassifikationen/Content75/KlassifikationWZ93.html, (21 April 2016).

Statistisches Bundesamt (2003): Klassifikation der Wirtschaftszweige, Ausgabe 2003 (WZ 2003), Wiesbaden. URL: https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaftklassifikationen/Content75/KlassifikationWZ2003.html, (21 April 2016).

Statistisches Bundesamt (2008): Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008), Wiesbaden. URL: https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaftklassifikationen/Content75/KlassifikationWZ08.html, (21 April 2016).

Toivanen, O. / Väänänen, L. (2012): Returns to Inventors. The Review of Economics and Statistics, 94 (4), 1173-1190.

Trajtenberg, M. (1990): A penny for your quotes: patent citations and the value of innovations. The Rand Journal of Economics, 21 (1), 172-187.

Trajtenberg, M. / Jaffe, A. / Henderson, R. (1997): University versus Corporate Patents: A Window on the Basicness of Invention, Economics of Innovation and New Technology, 5 (1), 19-50.

Ventura, S. L. / Nugent, R. / Fuchs, E. R.H. (2015): Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. Research Policy, 44 (9), 1672-1701.

Wermter, W. / Cramer, U. (1988): Wie hoch war der Beschäftigtenanstieg seit 1983? – Ein Diskussionsbeitrag aus der Sicht der Beschäftigtenstatistik der Bundesanstalt für Arbeit. Mitteilungen aus der Arbeitsmarkt – und Berufsforschung, 88 (4), 468-482.

Zedlitz, J. (2017): GOV – Das Geschichtliche Ortsverzeichnis, Computergenealogie 2/2017, 14-19.

6. Appendix

A1 Frequency tables

Frequency tables and overviews of the individual values and labels of the variables can be found in separate files at http://fdz.iab.de/en.aspx.

A2 Detailed description of data linkage

The INV-BIO ADIAB 8014 (henceforth: INV-BIO) data contain person-level information recorded in two large register data bases: the patent register of the European Patent Office (PATSTAT data) and the DPMA (DPMAregister) as well as administrative labor market data on employees and establishments available from the IAB. For the INV-BIO data, these register data have been linked at the person level, i.e. a link between inventors in the patent data and unique employees in the labor market data was established. The final data assigned inventor-patent records from the patent register to unique employees recorded in German social security data. The actual data linkage was performed in several steps as depicted in Figure A1.

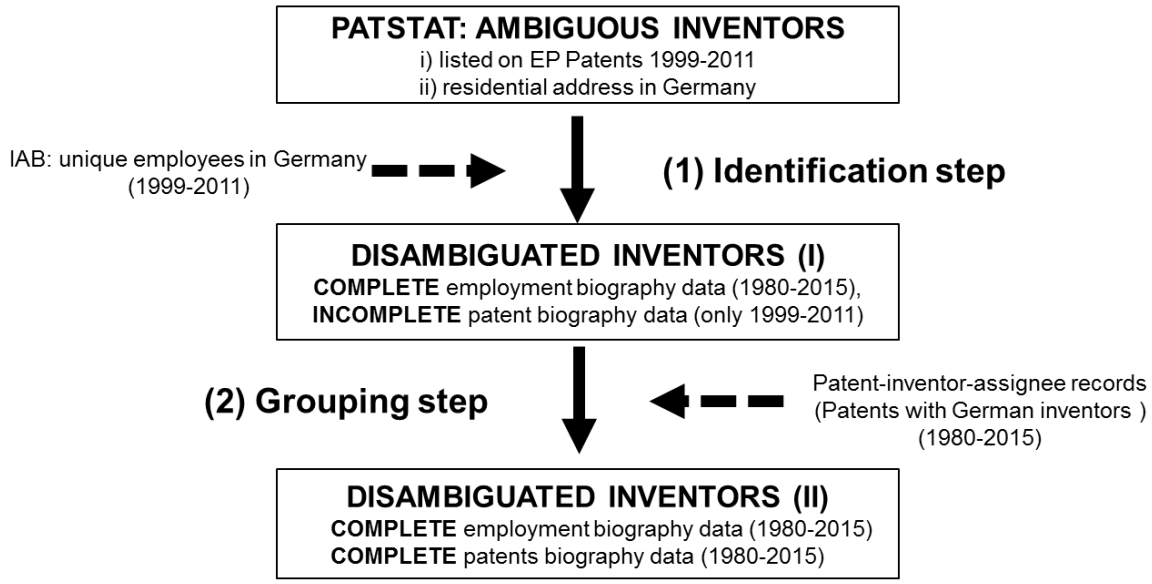


Figure A1: Data linkage workflow

1) Identification step

The main goal of Step 1 was the identification of inventors recorded on patents in the administrative labor market data of the IAB using methods of record linkage (Christen 2012). Data preparation involved at first the selection of all patent applications in PATSTAT (Version 10/2012) that include at least one inventor with a residential address in Germany (`person_country_code = "DE"`) during the period 1999 and 2011. Note that the period 1999 to 2011 was chosen because name and address data for the purpose of record linkage in the IAB are only available from 1999 onwards.

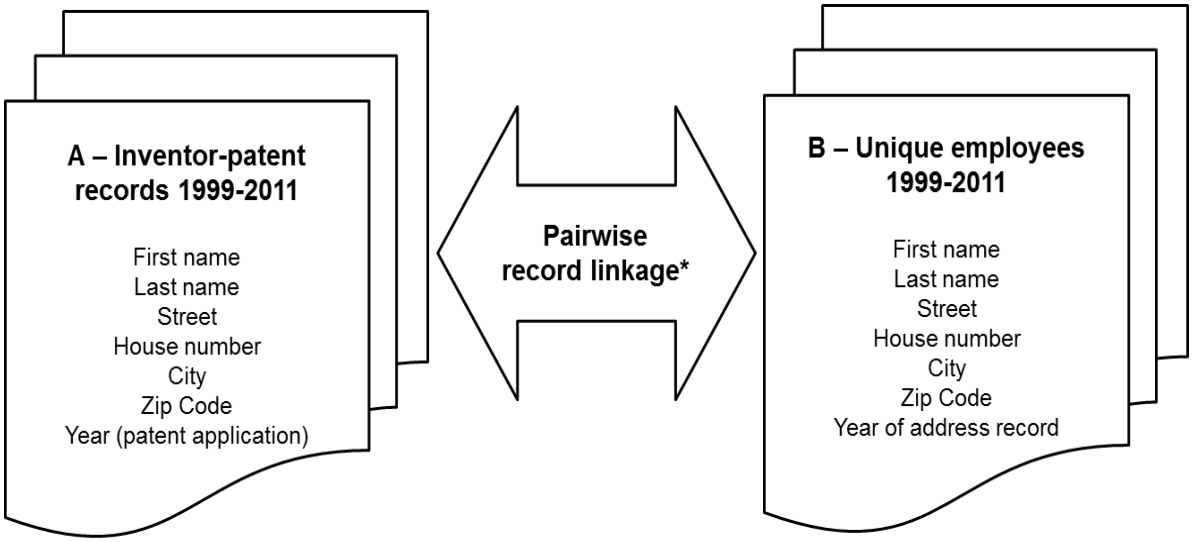
Name and address data on inventors in this data extract of PATSTAT were pre-processed (e.g. parsing of inventor names and addresses, standardization of umlaut and other special

characters, harmonization of city names based on zip codes, identification of academic degrees and honorifics) and prepared to match with the name and address data recorded on employees in the name and address data bases of IAB.[1] Prior to the linkage, the same string processing procedures were applied on the two raw data sets.

The comparison step of the record linkage process compared the pre-processed name and address information of persons in the two data bases in a sequence of linkage runs. First, a set of deterministic linkages was performed. The sequence of deterministic linkages varied the set of merging keys starting with a very restrictive version that defines matches conditional on name (first, last) and the full address information (street, house number, city, zip code) being exactly the same. In subsequent deterministic linkage runs, the conformity requirement of the full set of keys was gradually relaxed and record pairs were defined up to a setup in which the zip code in combination with the exact name match exactly across the source files.
The remaining set of unmatched records from the deterministic linkage were further processed using probabilistic record linkage techniques (Fellegi and Sunter 1969; Christen 2012, Chapter 6). These techniques relax the assumption of the merging keys to match exactly but tolerate some deviations (depending on the algorithm) that are due to different spellings or typographical errors in the input data. Eventually, probabilistic record linkage yields a similarity score that describes the univariate probability of a match across the sources. The open source software package we used, Merge Toolbox (Schnell et al. 2004)[2], features array matching functionalities that allow for computing similarity scores that combine weighted probabilities comprising multiple merging keys. In the final setup, the linkage algorithms (e.g. Jaro-Winkler algorithm for the names and n-gram algorithm for the address information[3]) and the weights (m-/u-probabilities)[4] for the individual merging keys were chosen following simulations performed at the German Record Linkage Center (Bachteler 2008). In the probabilistic record linkage runs, indexing metrics (also: blocking) were used to significantly reduce the number of string comparisons across the registers and, thus, computational requirements. These metrics reduced comparisons across the persons in the source files on data cells defined by the initial

[1] The raw adress data included also addresses on benefit recipients and job seekers, however these information were only used for the matching if information from the employment notification system was unavailable in the same year. Eventually, only a small fraction of addresses relating to non-employment records remained in the data set that was used in the subsequent record linkage steps.
[2] The Merge Toolbox (MTB) is an open source data linkage software package that was developed at the University of Duisburg-Essen and that is available from the website of the German Record Linkage Center at http://soz-159.uni-duisburg.de/software/index.html, last accessed 08 August 2018.
[3] For a detailed description of the record linkage algorithms see Christen (2012, Chapter 6).
[4] The m-probability is the likelihood of two keys matching if the records belong to the same individual. The u-probability is the likelihood of two keys matching if the records do not belong to the same individual. For further details on probabilistic record linkage, see Christen (2012, Chapter 6).

of the last name, year of the patent application with the EPO in combination with geographical characteristics derived from the zip codes in the inventor/employee addresses (see Figure A2).



* Record linkage setup: probabilistic string matching (array match of all merging keys), blocking by year x initials of last name x residential location.

Figure A2 Identification step and record linkage 1999-2011

Evaluation and post-processing involved the combination of deterministic and fuzzy links between patent-inventor records and individuals. Here, probabilistic matches were only considered if the similarity scores exceeded a cut-off threshold that was determined from a randomly selected training data set that was manually labeled.[5]

The consolidated linked data represent inventor biographies which contain inventor-patent records from the PATSTAT extract between 1999 and 2011 that are unambiguously assigned to a unique person in the IAB data.[6]

2) Grouping step

The main goal of the grouping step is to complete the inventor biographies that are limited to the period 1999-2011 after the first step (see Figure A1). Towards this objective, the full labor

[5] The training data was composed as a random sample of 100 probabilistic matches in equally sized bins of the similarity scores distribution. The actual threshold was defined at the minimum score of each bin, in which confusion matrix statistics precision and recall yielded satisfactory results. In the quality evaluation, a special emphasis was given to the minimization of the false positive rate, i.e. reduce the number of matches which have a high probability of being wrong, even at the cost of a lower number of potential true positive matches.

[6] The approach described here also solves the inventor disambiguation problem using an alternative methodology as typically used in the patent literature (e.g. Raffo and Lhuillery 2009, Li et al. 2014, Pezzoni et al. 2014, Ventura et al. 2015, Ge et al. 2016, Morrison et al. 2017). The work in this literature apply solely patent data and use sophisticated rules or machine learning techniques to cluster records with coherent characteristics, thereby obtaining a disambiguated inventor ID across patent records.

market biographies of the inventors as recorded in the Integrated Employment Biographies data base of the IAB (IEB) were extracted. For the actual implementation of the "grouping" approach, only employment records from these data were retained to generate an inventor-establishment panel data set structured by persons and employment episodes for years 1980 to 2014. These data include the name and the gender of the inventor, an imputed time consistent NACE industry classification[7] and location codes at the municipality level (8-digit code in AGS scheme) for the workplace and – from 1999 only – the residential address of the inventor.

From the PATSTAT register and the patent register of the German Patent and Trademark Office (DPMA), all patent applications that list at least one inventor with an address in Germany were extracted to construct an inventor-applicant panel data set that is equal to the structure described for the IAB data. Applications in these data that were filed with both the DPMA and the EPO were deduplicated and only the German applications (application authority "DE") were retained. Patents recorded in this patent register subset are the potential matches to complete the inventor biographies for which only links from 1999-2011 were made in the previous step. Prior to linking these data in the actual grouping application with the IAB data, extensive preprocessing of name, address and technology data was required.

First, applicants were labeled as corporations/organizations or individuals using a naïve Bayes classifier. For this, the preprocessed names (and legal forms) in the OECD HAN data base were extrapolated to the DPMA data.[8] Individuals who are recorded as patent applicants were discarded afterwards because address information of individuals hardly compares to actual establishment locations in the subsequent steps of the grouping.

Second, preprocessing of address information on inventors and the remaining (corporate) applicants involved the translation of 4-digit zip codes (used until 1993) into the current 5-digit zip code system.[9] In the subsequent step, inventor and applicant address data were geocoded using an offline geocoding tool licensed by the IAB. These geocoded inventor and applicant address data were enhanced and cross-checked with geocoded data provided by the Max Planck Institute for Innovation and Competition.

[7] The time consistent NACE Rev. 1 industry classification was implemented based on the methodology described in Eberle et al. (2011).

[8] The OECD HAN data base includes inventor and applicant names after comprehensive cleaning routines. Based on the applicant names, the data also propose a sector allocation. The OECD HAN data can be downloaded from OECD website: http://www.oecd.org/sti/intellectual-property-statistics-and-analysis.htm#ipdata (last access 2018-07-23).

[9] For the translation of zip codes, online resources such as data from the Geschichtliches Ortsverzeichnis (GOV) (Zedlitz 2017) available at http://gov.genealogy.net and http://www.alte-postleitzahlen.de were used.

Third, detailed IPC codes describing the technology of the patent application were consolidated into 34 aggregate technology areas following the WIPO IPC Technology Concordance Table dating back to Schmoch (2008).
Based on these preprocessed data, an inventor-applicant panel with population data on patents and inventors from 1980 to 2014 was generated.

The actual grouping of inventor-applicant panel data (B file) to the previously identified inventor accounts (inventor-establishment panel data, A file) proceeded in two steps and is displayed in Figure A3. The first step involves the generation of relational features and the definition of a candidate set with record linkage (visualized as the arrows). In the second step, a machine learning work flow is implemented that uses supervised classification methods to assign and group patent records to unique individuals.

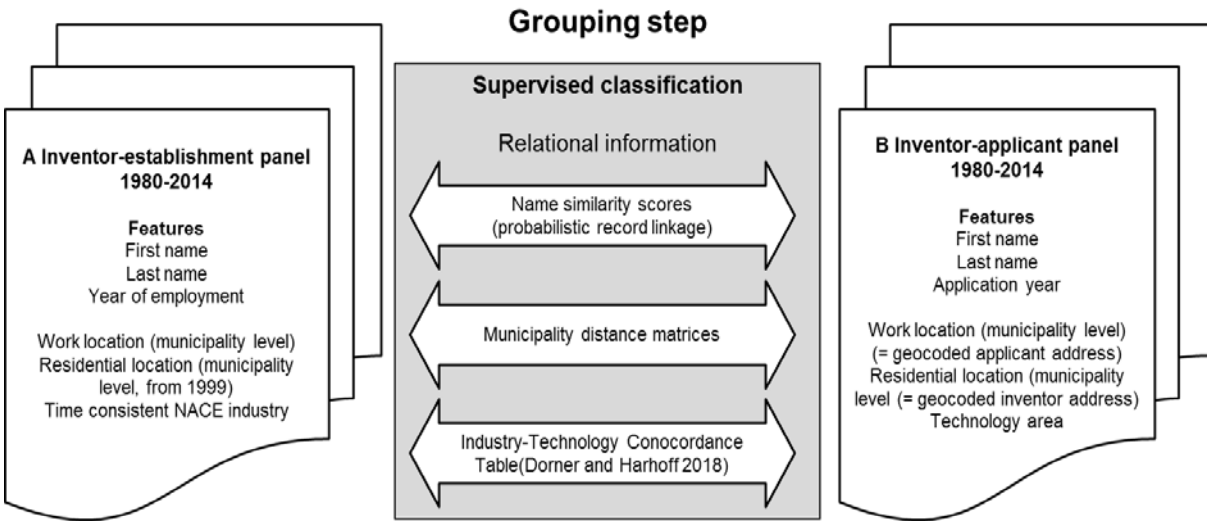


Figure A3 Grouping step

First, name similarity scores generated from a (probabilistic) record linkage on the inventor names were used to identify potential candidates with similar names for each inventor establishment record in the inventor-establishment file. Again, blocking based on data cells that are defined by the Soundex codes, i.e. a phonetic encoding of the inventor last names and years were used to enhance computational efficiency of the linkage.
After implementing an empirical threshold that distinguishes potential matches from non-matches based on the name similarity score, the resulting data set includes for all employment records of the inventors links to patents in the same year that are filed by inventors with the same or similar names.

For these potential/likely matches, in another feature engineering step, further relational variables were produced that facilitate the comparison between the reference and the candidate record. These features (see Figure A3) include additional name-based similarity variables (e.g. binary indicators for the same name components), an indicator for the same gender and geographical distance matrices for residential and work locations, respectively. Additionally, the industry-technology concordance matrix by Dorner and Harhoff (2018) was merged to describe the likelihood of filing a patent in technology *t* while the inventor is being employed in industry *i*.

In the second step, a (supervised) machine learning workflow was developed to filter false positive matches from the candidate sets that were previously generated with the name-based probabilistic record linkage. The underlying machine learning problem can be described as a classification task that was addressed using algorithms typically used in supervised classification. The implementation proceeded as follows: A random subsample of the linkage results comprising 3,200 records that link inventor-patent records to employment episodes of similar inventors were extracted as training data.[10] These records were manually classified as matches and non-matches evaluating the comprehensive set of relational features within a set of potential/likely matches. For instance, it is possible that a probabilistic link across the sources was generated solely based on a common German name (e.g. Andreas Schulz). However, the additional relational features could point to an unlikely match because the linked records refer to distant locations and/or the technology of the patent does only in very rare cases originate in the industry the inventor is actually employed in. Employing the labeled (training) data, a logistic classification model was set up to predict the matches in both the training and in the test data (out-of-sample prediction).[11]

We use repeated 2-fold cross validation on the labeled training data to select the specification of the actual predictive model.[12] The cross validation divides the training data into two halves, with one half being used as a validation data set in which the manually assigned labels are compared to the predictions from the classification model that was fitted on the remainder of the data. This procedure is repeated 500 times. The comparison of the results in each of the

[10] Using with cross validation, we evaluated based on precision and recall metrics with a gradually increasing number of training records (starting from 100 up to 3,200) whether the number of 3,200 manually labelled records is sufficiently large for valid predictions. Results of the cross validation analysis for the preferred setup indicated that after passing a threshold of approx. 700 records in the evaluation and hold-out set each, did not change the results of the analysis.

[11] For the use of logistic regression in settings of supervised classification, see Hastie et al. (2009, Chapter 4).

[12] For further information on cross validation techniques, see Hastie et al. (2009, Chapter 7).

500 cross validation runs can be tabulated as a confusion matrix, which forms the basis for the computation of more comprehensive metrics such as precision and recall. The term precision denotes the ratio of correctly predicted positive observations (matches) to the total predicted positive observations. Recall describes the ratio of correctly predicted positive observations to the all observations in actual class of correct matches (including true negatives). These indicators are frequently used to assess the quality of the prediction using alternative views on cost and utility (Christen 2012, Chapter 7). Thus, these measures describe the trade-off between a restrictive model that minimizes false positive records on the one hand and a rather relaxed model variant that optimizes the number of matches at the expense of having more potentially false positive records in the results. Precision and recall can be aggregated using the harmonic mean, which is then called the F-score (Christen 2012, Chapter 7). By is mathematical definition, F-score has only a high value if both the precision and the recall metrics reach high values. Thus, optimizing the F-score yields the best compromise between precision and recall.

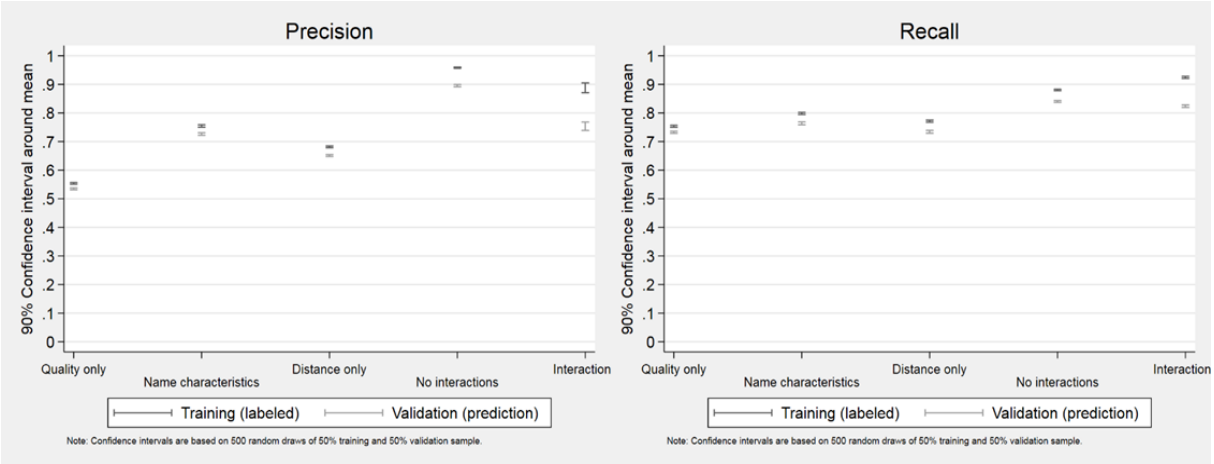


Figure A4 Cross validation results on model specification

To identify the optimal classification model, we tested various specifications in the cross validation (see Figure A4; different specifications on the x-axis). We used the maximum F-score as the selection criterion to determine the best model specification. We start with three basic models that include only one dimension each describing the similarity score of the record in A and B file. In the fourth specification, termed “No interactions”, we combine the rich set of name similarity features, distance features and the industry-technology concordance information to predict the probability of an actual match. The fifth specification (“Interaction”) includes the same features plus a set of additional polynomials of the variables to model the relationship of records in A and B file with more flexibility. The precision and recall estimates

of the 500 cross validation runs depicted in Figure A4 show that the “No interaction” specification outperforms the other models. The drop in precision in the “Interaction” specification points to issues with overfitting of the model, while the results are similar for recall. As expected, the one dimensional prediction models do not reach the predictive quality of the more comprehensive specifications.

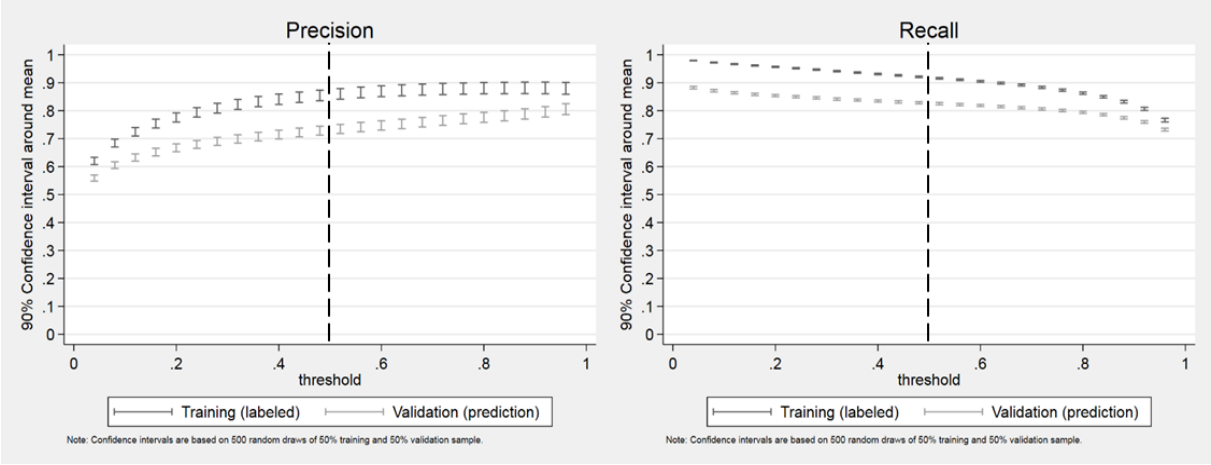


Figure A5 Cross validation results on cutoff threshold

In a logistic classification model a cutoff value for the predicted probability of the record being a match is required. This threshold was determined using a grid search with the maximum F-score as the selection criterion. Here, we explicitly show this process for the preferred model specification “No Interactions”. We tested the impact of varying the cut-off value for the predicted probabilities between zero and one to determine the actual matches. We find that the default threshold of .5 to classify records in matches and non-matches appears to yield a good compromise between precision and recall.

The F-Scores, which summarize the optimal compromise between precision and recall, are displayed in Figure A6. The left panel of Figure A6 shows the specification grid search analysis (see also Figure A4) and the right panel displays the search for the optimal cut-off threshold of the logistic classifier (see also Figure A5).

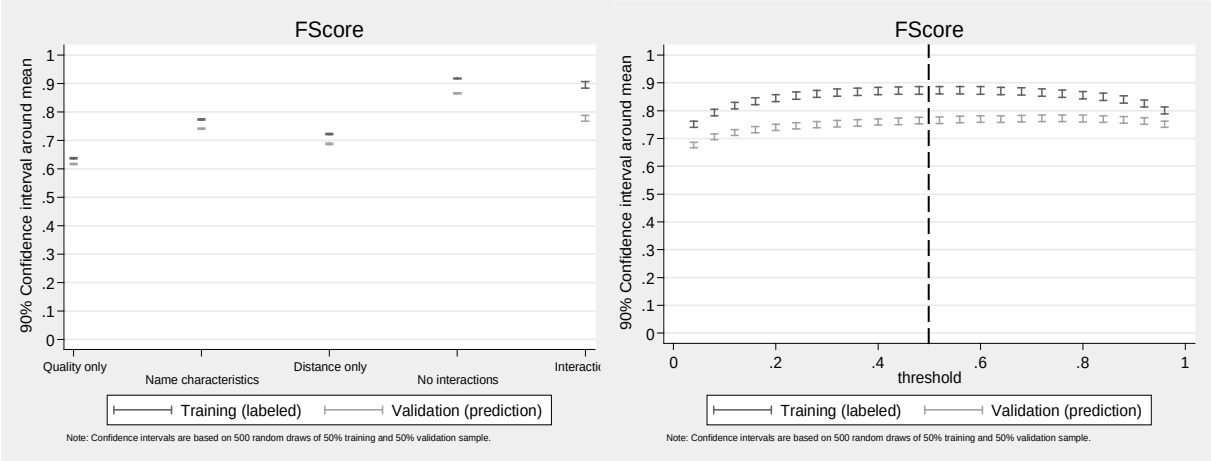


Figure A6 F-Scores of the cross validation analysis

Using the optimized model setup for supervised classification, the training set error (misclassification error rate) of the final model amounted to 1.75 percent. Employing alternative classification methods such as variants of discriminant analyses yielded inferior matching quality compared to the logistic classifier.[13]

Since there is no information available whether the label predicted by the model based on the randomly selected training sample is actually correct for the records in the test data, the corresponding test set error cannot be calculated.

[13] For a comparison of discriminant analysis and logistic regression in classification settings, see Hastie et al. (2009, Chapter 4).

A3 Representativity of the data evaluated against PATSTAT

We do not observe the number of disambiguated inventors in Germany for the period 1999 until 2011. Therefore, we are not able to evaluate our data linkage at the level of individual inventors. Representativity, however, can be assessed at the level of patents, which gives a reasonable good approximation of the linkage quality.

We compare the number of DOCDB patent families – reference data point is the earliest application filing date within DOCDB family – linked to unique inventors in our data with corresponding patent families recorded in the full EPO (PATSTAT data) and DPMAregister, respectively. We distinguish between the full population of all patents in the register and a subset of likely German patents. The latter are defined as patents that list at least one inventor and applicant with the country code "DE", as assigned by the patent office based on address information reported on the patent forms. We assume that these German patents originate in some inventive activities performed in the German national innovation system, of which the labor market is considered a fundamentally important subsystem.

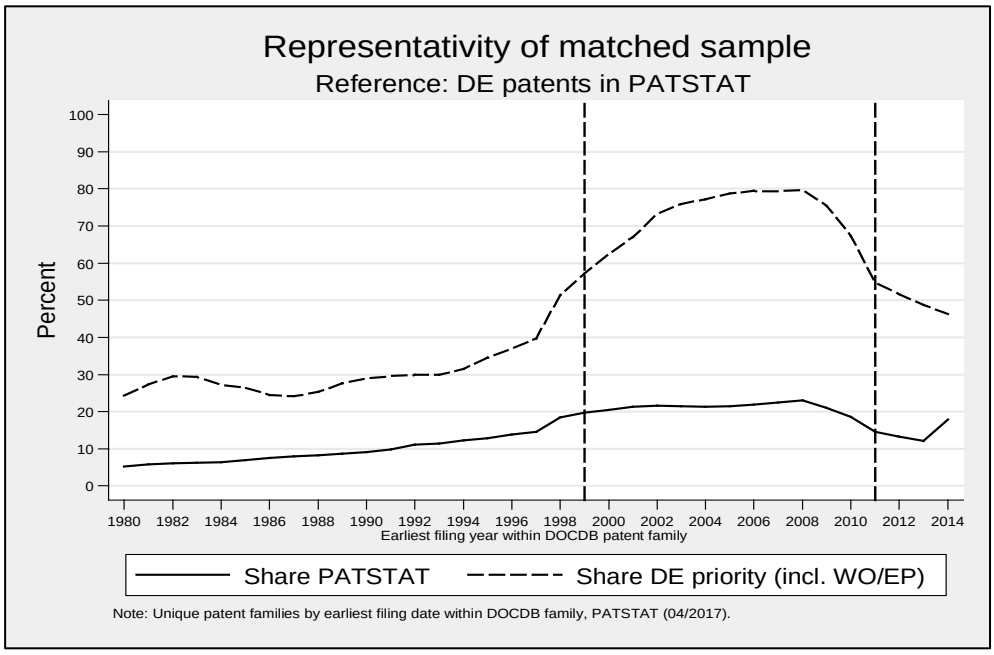


Figure A6 Representativity at the level of patent families

Figure A6 depicts the annual ratios of patents with at least one linked inventor being represented in our sample at the level of patent families. For the German patents, in the period 1999 to 2011, our sample represents on average about 71.4 percent of the DOCDB patent families in the population. The maximum coverage of almost 80 percent is realized just before the global financial crisis starting in 2008. As a result of the cohort approach, the representativity drops outside of the 1999-2011 time window to a level of about 30 percent of

the patents in Germany that are comprised in the sample. For the full population of patents in the PATSTAT and DPMAregister, representativity reaches levels of about 20 percent in the 1999-2011 time window and levels of 10 percent otherwise.

Patents originate in very different technological regimes and industrial sectors (see also Dorner and Harhoff 2018). While some segments of the market of technology are very dynamic and require up to date scientific knowledge, skills and cutting edge work environments or tools, others are rather persistent with a relatively long half-life of field specific knowledge and/or low requirements for complementary tools. To shed light into this source of variation in the data, representativity is further assessed across the five technology main areas that are defined in the broad technology classification of the WIPO (Schmoch 2008). According to this approach, a patent is classified by its modal technology field (accounting for the full IPC information available) and subsequently categorized into one of the five unique technology main areas. The curves in Figure A7 visualize the ratio of German patents by main area that are represented in the inventor-patent data of the INV-BIO data set. In line with the population analysis, the same temporal pattern emerges in the evaluation by technologies.

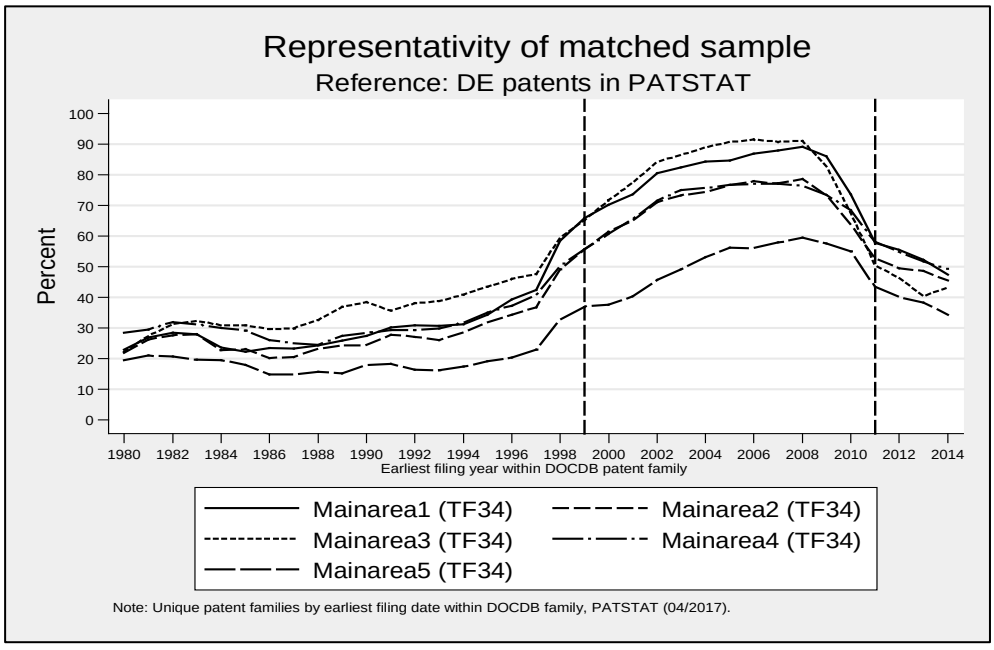


Figure A7 Representativity at the level of patent families by technology main area

Further, Figure A7 shows that across main areas, there is substantial variation to the extent patents are represented in the INV-BIO data. The best representation of patents in the 1999-2011 time window is achieved for the main area 3 *Chemistry*. In some filing years, even more than 90 percent of the inventions in Germany assigned by the patent examiners to technology

field are represented by matched inventors in the INV-BIO data. On average, the representation remains very high at a level of more than 80 percent. Also, the population of *Electrical engineering* patents (main area 1) is largely represented in the INV-BIO data. Fields such as *Instruments* (main area 2) and *Mechanical engineering* (main area 4) show similar representation rates of about 70 percent on average in 1999-2011 observation window. The representation of patents assigned to the main area 5 *Other Fields*, however, drops substantially as compared to the ones discussed before. One reason for the lowest representation rate of this technology field is that a large part of these patents are filed in civil engineering, an industrial segment in which public sector civil servants or researchers as well as self-employed architects presumably play an important role. Both groups are not recorded in the social security data on employees that were used to link inventors.

Overall, given that our matching is entirely based on individual level characteristics of inventors or employees, respectively, we argue that the population of patents and individuals represented in the INV-BIO is a random population sample.
Given the large size of the sample in the 1999-2011 time window, we argue that estimates obtained from the data give already a very good approximation to population estimates, although, due to the lack of information on the actual inventor population, no inventor level weights can be provided for statistical projection to the population.

A4 Stata syntax for merges across the INV-BIO files

Most applications with INV-BIO data will require the researcher to link files of the INV-BIO data set with each other. An overview on the merging keys is given in Section 1.3. In the following, examples are discussed and syntax to be used with Stata (here: version 14) is documented.

First, when linking the labor market biography file with the inventor patent file, it has to be taken into account that different research goals will require different data structures. For instance, if the biographies of the inventors and the actual trajectories are of interest, then the two files may be merged using a full join (Stata command `joinby`).

Pre-processing is required for linking the INV-SIAB and INV-PAT data, because the latter includes only quarterly data. The pre-processing involves the extraction of the year of the patent filing, which is required for the join with the INV-SIAB employment episodes.

```
*** STEP #1: INV-PAT pre-processing

* Define set of patent variables to be used for the example analysis
* Note: if required, additional variables should be added.
global VAR_PAT_INV_INPUT "docdb_family_id appln_id earliest_filing_date
appln_filing_date nb_inventors nb_applicants area34"

* INV-PAT data
de using "orig\INV-BIO_8014_v1_INV-PAT.dta",s
use erf_id $VAR_PAT_INV_INPUT using " orig\INV-BIO_8014_v1_INV-PAT.dta", clear

* Year of earliest filing date (efd)
gen int yr_efd = yofd(dofq(earliest_filing_date))

save "data\tmp_overlap.dta", replace
```

Box A4-1: Example code for Stata

After the pre-processing of the patent records, some pre-processing of the dates for the following steps is also required in the INV-SIAB. Besides that, the pre-processing involves the creation of the employment duration variable and the identification of the main employment of each episode.

```
*** STEP #2: Combine INV-SIAB data with patent information
de using "orig\INV-BIO_8014_v1_INV-SIAB.dta", s
use "orig\INV-BIO_8014_v1_INV-SIAB.dta" if quelle == 1, clear

* Tenure in establishment at the start of the employment episode
* Note: Considers all employment episodes for simplicity, regardless of the type
* of employment
```

```
sort erf_id betnr spell
gen int durepi = (endepi-begepi)+1
by erf_id betnr: gen long dur_est = sum(durepi)
drop durepi

* Identify main employment episode (among potentially simultaneous episodes) using
* source, employment status and wage and drop side jobs
gsort erf_id begepi endepi erwstat -tentgelt
by erf_id begepi endepi: keep if _n == 1

* Year variable derived from begepi for join
gen int yr_efd = year(begepi)

* Full join by inventor and year
de using "data\tmp_overlap.dta"
joinby erf_id yr_efd using "data\tmp_overlap.dta", unmatched(both) _merge(_merge)
tab _merge
```

Box A4-2: Example code for Stata

At the level of episodes, this join yields three types of results:

1. Exact 1:1 matches of episodes, which indicate the focal inventor filed one patent during one employment episode and which overlaps the filing date.
2. m:1 links of patents and employment episodes, i.e. multiple patents are assigned to one unique episode and the initial record is multiplied by the number of linked patents. Please note, if subsequent analyses are performed in this duplicated data structure, researchers might want to aggregate or prioritize patent filings for the same employment episode (e.g. generate a (citation weighted) count of patents by employment episode).
3. m:n links of patents and employment episodes resulting from the full join on the inventor ID and the year of the patent filing. Here, the same focal invent-patent record is matched with more than one employment episode of the inventor in the same year. This special case is only possible if the same inventor has patent filings and multiple employment episodes in the same year. To identify and solve m:n relationships, the indicator variable `overlap` is generated. This variable marks records with the value of one, in which the patent filing date lies within an employment episode. Duplicated records with `overlap == 0` may be discarded.

```
* Indicator variable for patents that are filed during employment episodes
sort erf_id spell
```

```
gen byte overlap = earliest_filing_date >= qofd(begepi) & earliest_filing_date <=
qofd(endepi)

* Check how many patents are filed within employment episodes
tab overlap quelle

* Recode patent variables to NA for non-overlapping employment episodes
foreach var of global VAR_PAT_INV_OUTPUT {
       replace `var' = . if overlap == 0
}
```

Box A4-3: Example code for Stata

For analyses at the level of patents that focus on the characteristics of the inventors or establishments as recorded in the INV-SIAB or INV-BHP data respectively, the data may be reduced to the employment episodes in the data in which patents are filed.

```
* Keep only patent records that are related to employment episodes
tab quelle overlap
keep if overlap == 1 & quelle == 1
drop overlap
```

Box A4-4: Example code for Stata

For inventors who have multiple employment episodes with different employers (establishments) in a quarter in which a patent was filed, the data records duplicates (representing the number of employment episodes). To clear these duplicates and unambiguously assign the patent filing to an employment episode and employing establishment, the researcher might want to assume that the episode in which the inventor has the longest tenure is the plausible record to retain and reference to the patent.

```
* Sort duplicates descending by establishment tenure to clear duplicates
gsort erf_id docdb_family_id -dur_est
by erf_id docdb_family_id: keep if _n == 1
isid erf_id docdb_family_id
by erf_id docdb_family_id betnr: assert _N == 1
```

Box A4-5: Example code for Stata

Alternatively, also the episode with oldest/youngest date can be used as reference. The sorting changes accordingly, nevertheless, the actual employment episode may change, which may affect also results of subsequent analyses.

The resulting data set after this preparation is restricted to unique inventor-patent (docdb_family_id) records for which inventor and establishment information are potentially available in the administrative labor market data of the IAB. In this data structure, for instance, the structure of inventor teams may be analyzed.

Alternative approaches for data analyses are possible, e.g. by not restricting the data set to episodes with patent filings (`overlap = 1`), but instead using the biographical perspective to study the career trajectories of the inventors following the structure of the administrative labor market data.

A5 List of abbreviations

AA	Agentur für Arbeit / Arbeitsamt	Employment agency / employment office
AGS	Amtlicher Gemeindeschlüssel	Official Id for administrative units
ALG	Arbeitslosengeld	unemployment benefit
ARGE	Arbeitsgemeinschaft	cooperation of employment agencies and municipalities
ASU	Arbeitsuchende-Historik	Jobseeker History
A2LL	Arbeitslosengeld II – Leistungen zum Lebensunterhalt	unemployment benefit II - benefits to secure a livelihood
BA	Bundesagentur für Arbeit	Federal Employment Agency
BeH	Beschäftigten-Historik	Employee History
BHP	Betriebs-Historik-Panel	Establishment History Panel
coArb	Computerunterstützte Arbeitsvermittlung (operatives Verfahren zur Verwaltung der Vermittlung (Altverfahren))	computer-aided job placement (procedure for the administration of job placements – old procedure)
DEÜV	Verordnung über die Erfassung und Übermittlung von Daten für die Träger der Sozialversicherung – Datenerfassungs- und Datenübermittlungsverordnung	Data Collection and Transmission Regulation - regulation on the collection and transmission of data for the social security agencies
DEVO	Zweite VO über die Erfassung von Daten für die Träger der Sozialversicherung und für die BA – Datenerfassungs-Verordnung –	Data Collection Regulation - second regulation on the collection of data for the social security agencies and for the Federal Employment Agency
DPMA	Deutsches Patent- und Markenamt	German Patent and Trademark Office
DÜVO	Zweite VO über die Datenübermittlung auf maschinell verwertbaren Datenträgern im Bereich der Sozialversicherung und der BA – Datenübermittlungs-Verordnung –	Data Transmission Regulation - second regulation on the transfer of data on machine-readable data media in the field of social security and the BA
EPO	Europäisches Patentamt	European Patent Office
FDZ	Forschungsdatenzentrum	Research Data Centre
FELEG	Gesetz zur Förderung der Einstellung der landwirtschaftlichen Erwerbstätigkeit	Act on the Support in Case of Termination of Farming Activities
IAB	Institut für Arbeitsmarkt- und Berufsforschung	Institute for Employment Research
IEB	Integrierte Erwerbsbiographien	Integrated Employment Biographies
ISIC	International Standard Industrial Classification of All Economic Activities	International Standard Industrial Classification of All Economic Activities
KldB	Klassifikation der Berufe	Classification of Occupations
LeH	Leistungsempfänger-Historik	Benefit Recipient History

LHG	Leistungs-Historik Grundsicherung	Unemployment Benefit II Recipient History
MTH	Maßnahmeteilnehmer-Historik	Participants-in-Measures History File
NACE	Nomenclature génerale des activités économiques dans les communautés européennes	Nomenclature génerale des activités économiques dans les communautés européennes
NUTS	Nomenclature des unités territoriales statistiques	Nomenclature des unités territoriales statistiques
SGB	Sozialgesetzbuch	German Social Code
SIAB	Stichprobe der Integrierten Arbeitsmarktbiographien	Sample of Integrated Labor Market Biographies
VerBIS	Vermittlungs- und Beratungsinformationssystems	Information System for Placement and Counselling
WIPO	Weltorganisation für geistiges Eigentum	World Intellectual Property Organization
XASU	Arbeitsuchenden-Historik aus XSozial-BA-SGB II	Jobseeker History from XSozial-BA-SGB II

Imprint

FDZ-Datenreport 3/2018

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2018/DR_03-18_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Dr. Matthias Dorner
Institute for Employment Research (IAB)
Phone: +49-911-179-1752
Email: matthias.dorner@iab.de

Prof. Dietmar Harhoff, PhD
Max Planck Institute for Innovation and
Competition
Email: dietmar.harhoff@ip.mpg.de