DRUID Working Paper No. 10-16

The Danish Integrated Database for Labor Market Research: Towards Demystification for the English Speaking Audience

By

Bram Timmermans

Danish Research Unit for Industrial Dynamics

www.druid.dk

DRUID

The Danish Integrated Database for Labor Market Research: Towards Demystification for the English Speaking Audience

Bram Timmermans
Aalborg University
Fibigerstræde 4
DK-9220 Aalborg Ø
Denmark
E-mail: bram@business.aau.dk

February 17, 2010

Abstract:
An increasing number of studies, in a wide range of disciplines, make use of so-called linked employer-employee databases. These detailed databases are only available in a limited number of countries. Denmark is one of the countries that makes this data available to researchers that are connected to a Danish research institute. Due to the sensitivity of the data, access is only granted after a thorough screening process and under strict conditions. Nevertheless, there is an international interest in the structure of this database and what information it provides. The purpose of this document is to provide such an description in English.

Keywords:

Jel codes:

ISBN 978- 87-7873-298-9

1 Introduction

Determining the impact of human resource decisions made by firms requires detailed information on these human resources. This information needs to be obtained through elaborative data gathering. Earlier-conducted studies that take this human resource perspective traditionally implement a qualitative approach, using case studies and surveys. Despite that these approaches can provide in-depth information on a specific set of cases, they are expensive and time consuming to undertake. As a result, they focus on a select number of firms and often target a smaller set of individuals within these firms (e.g. top management teams). Nowadays, the access to linked employer-employee databases can solve parts of this data-gathering problem. These databases enable me and other researchers to (i) select a large sample of firms, including their entire work force, and (ii) a large amount of objective variables describing both the characteristics of the firm and the individuals attached to these firms. Despite the fact that the data contains no direct information on decision making processes it does contain information on the composition of workers and worker flows that can be associated with firm founding, growth and disbanding (Eriksson and Kuhn, 2006). Databases similar to the one used in this dissertation will provide a wide range of researchers with the opportunity to test a variety of theories and examine the generalizability of more qualitative approaches (Campbell, 2005). In addition to this firm perspective, these databases can be used to study different phenomena occurring on a more aggregated level (e.g. industry life cycles).

The linked employer-employee database described in this document is the Danish Integrated Database for Labor Market Research - in Danish ”Integreret Database of Arbejdsmarkedsforskning”.[1] This database is used in a wide range of scientific disciplines, varying from labor market economics to health sciences. Most of these researchers provide a description of the database similar to following quote by Sørensen (2004):

> ”To investigate these issues, I take advantage of unusually rich and comprehensive data on the dynamics of the Danish economy. Because its welfare state policies (and tax demands) are so extensive, the Danish government keeps unusually comprehensive information, by U.S. standers, on the economic activities of firms and individuals. The underlying datasets for these analyses come from databases maintained by Statistics Denmark under the name 'Integreret Database for Arbejdsmarkedsforskning (IDA)' (Integrated Database for Labor Market Research'). IDA combines information on individuals and establishments from a variety of registers maintained by the Danish government.
>
> There are several important elements that characterize the data from IDA. First, IDA contains rich information on individual characteristics, including

[1] Similar databases are, as far as I have been informed, available in Sweden, Finland, Norway, and Finland. Other large sets of micro-data can also be found in Germany and the United States.

> information on education (length and type), work experience, wages and income, wealth, unemployment, sex and age. Second, individuals in the labor force are matched to establishments and employers, which can be characterized in various ways, including industry affiliation. Third, the data are longitudinal, being updated annually since 1980. This means that people who change employers can be tracked." (Sørensen, 2004, p. 155-156)

It is understandable that researchers limit the description of this database to a couple of paragraphs. Nevertheless, I will take the opportunity to go more into detail in describing the characteristics, the opportunities, and the limitation of the Danish Integrated Database for Labor Market Research (from now on referred to by its Danish acronym, IDA). In doing so I will present (i) the structure of the database, (ii) the different variables listed in the database, and (iii) the limitations of the database.

The main motivation writing this document is that throughout my career as a doctoral student, many questions have risen on the characteristics of this databases. I could refer to two sources that present a more elaborate description of the database. First, the website of Statistics Denmark[2] with limited English documentation on the database. Second, a document with a thorough description of IDA created in 1991 by the project group that was in charge of creating the database (Emerek et al., 1991). Unfortunately, this description is only available in Danish.

These two sources are the foundation of the description of IDA presented in this document. The variable description will be based on the IDA database that is available to the researchers of the IKE research group at the Department of Business Studies, Aalborg University.[3] A longer list of variables can be found on the website of the Aarhus School of Business.[4]

Furthermore, I will describe two other databases that will be merged together with IDA and are used in the analyses in the chapters that follow. First, the database containing firm level statistics, mostly accounting statistics, also maintained by Statistics Denmark. Second, the DISKO4 survey on organizations, employees, and research and development strategies in Danish firms. This innovation survey was sent to a stratified sample of

[2] www.dst.dk

[3] Access to IDA is only given to research environments that can be considered permanent. Those environments that can be approved are: publicly funded research projects, employees at public research environments, employees at charitable foundations. Private parties are restricted from access but in special circumstances can get access. Foreign researchers can only get access when they are connected to an authorized Danish research environment. Access is made possible after a thorough screening process and against payment to Statistics Denmark after which the data is made anonymous, i.e. personal and establishment data receives a random identification number. In addition, Statistics Denmark has a set of rules in order to maintain discretion and data security.

[4] Documentation can be found on http://www.asb.dk/article.aspx?pid=675. Alternatively, go to their website (www.asb.dk) and search for "IDA variables".

Danish firms with more than 20 employees in 2006.

2 A Description of the Integrated Database for Labor Market Research

The project group in charge of creating IDA consisted out of researchers from Aalborg University (Ruth Emerek), Copenhagen Business School (Per Vejrup Hansen) and Statistics Denmark (Søren Leth- Sørensen). Over a three-year period, 1988-1990, they developed the database using government registers from 1980 and onwards. This database has been updated ever since on an annual basis. As a result, researchers in a number of different disciplines have access to data that (i) connects individuals with establishments and firms, (ii) is longitudinal, and (iii)is universal. Furthermore, the fact that the data is created by researchers to conduct research is a strong feature of this database. Figure 1 shows the base structure of this IDA database.

As this figure illustrates, IDA consist out of three type of databases that are integrated along two dimensions (i.e. vertical and horizontal). On the vertical dimension, a merge is possible on the personal, employee, and establishment level for any given year. The characteristics of these three levels will be described in the following sections. The information in these databases is obtained at one point at any given year. The link between individuals and establishments is identified at the end of November. This period is chosen since it is close to the end of the year and there is less interference with seasonal effects (e.g. the closure of firms around Christmas). Because the employment relationship is measured at one point during a year; as a result, seasonal jobs and multiple job changes within a year are not observed. Personal variables are predominantly taken at the end of the year; exceptions are the education variables, which are obtained in October. The horizontal dimension presents a longitudinal perspective. Creating the possibility to observe changes in the labor market from 1980 and onwards, covering more than 25 years.[5] (Emerek et al., 1991)

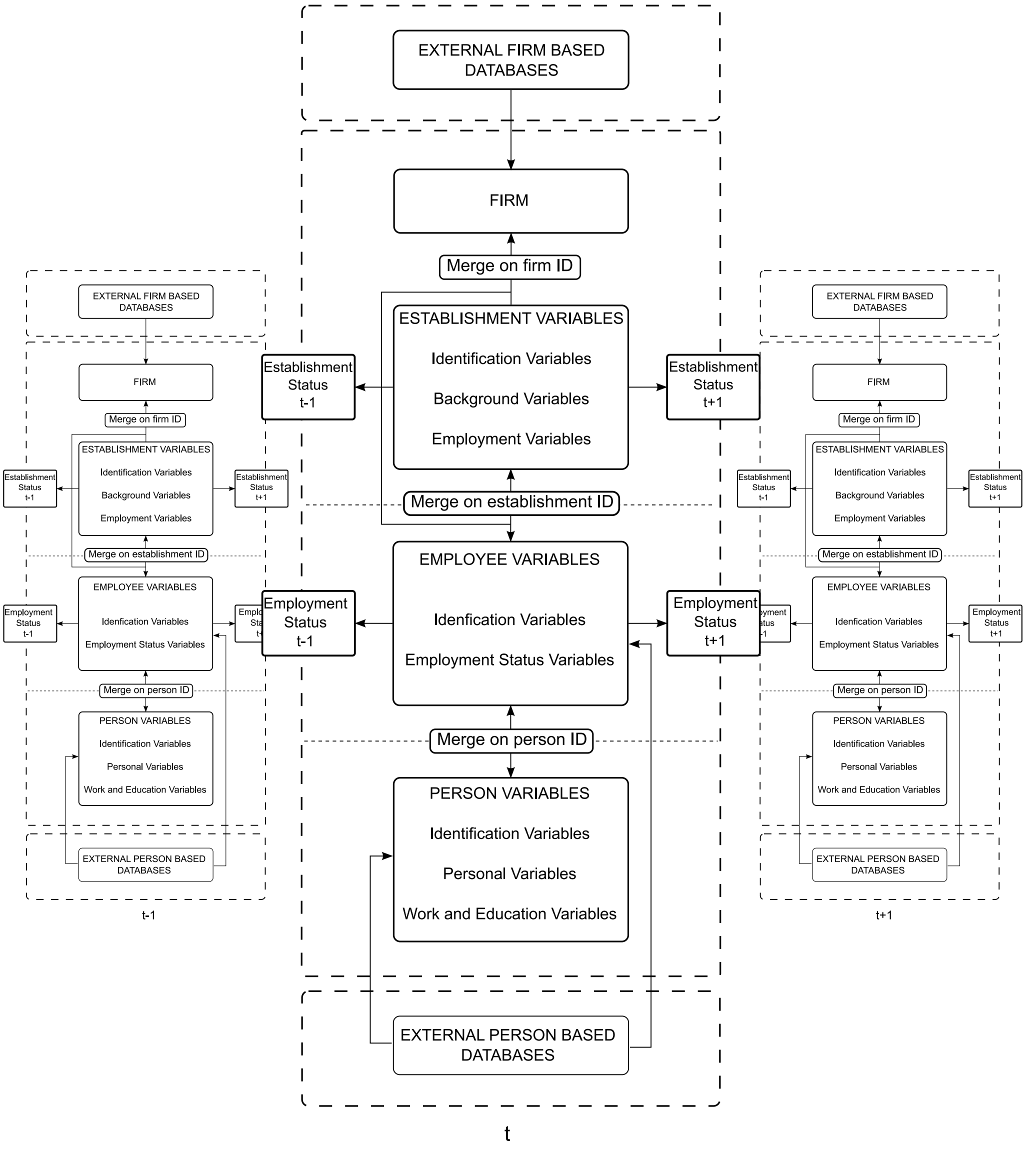
2.1 Establishment level

An establishment[6] is an organizational unit with its own type of activities and/or geographical location. Consequently, one firm can consist out of multiple establishments. The number of establishments varies between 172,000 and 187,000, divided over approximately 134,000 to 150,000 firms in the period 1980 to 2006. The number of firms with multiple establishments is, however, limited. Only 3.5 to 4.6 percent of all the listed firms have more than one establishment. The detailed establishment information can be

[5] In this case the data is available until 2006. In practice, data is available up to two years before the current year.

[6] In this document I use the term ”establishment”; in many other studies you may encounter the term ”plant” instead. These two terms can be used interchangeably.

Figure 1: Base Structure of IDA



categorized in three different groups (i.e. identification variables, status variables, and retrospective and prospective variables). A selection of variables that will be used in the different analyses are listed in Table 1.

Each establishment has two *identification variables*. First, a unique ten-digit establishment identification number. The first four digits are equal to the year in which the establishment appeared for the first time in the November registration period[7], followed

[7] If an establishment already existed prior to the start of IDA (i.e. 1980) the establishment is registered

by a random six-digit number. The second identification number identifies firms by providing a so-called employer identification number. With this number it is possible to identify those establishments that are part of the same firm.

One strength of IDA is the time perspective of the establishment identification number. Over time, an establishment can go through a number of changes (e.g. change of ownership, change in industry, change of location, change of labor force, change of products, etc). The question is whenever such a change would mean a change in identity. For example, it would be incorrect to treat a change in ownership, which results in a change of the firm identification number, as a change in establishment identity if all the other characteristics remain the same. IDA follows, due to the wish to follow establishments over a longer period in time and to track labor mobility patterns, a person oriented approach towards change. This means that an establishments identification number stays the same from one year to the other whenever one of the following criteria is fulfilled:

1. an establishment has the same owner and is active the same industry;
2. an establishment has the same owner and the same labor force; or
3. an establishment has the same labor force and is located on the same address or is active in the same industry.

There is a strict definition to what is meant with the same owner, same industry, same address, and same labor force. An establishment has the same owner whenever the firm identification number is the same from one year to the other. Being active in the same industry means that the four-digit NACE industry code remains the same between the two consecutive years. The same address refers to the same municipality and street code. The definition of the same labor force varies between lines two and three. In line two, at least 30 percent of the labor force should be present in one of the two years. Thus, either 30 percent of the employees that worked in year 1 should be present in year 2 *or* 30 percent of the employees in year 2 should have been present in year 1. In the third line, when there is a change in ownership, at least 30 percent of the employees should be present in both years. This means that 30 percent of the employees that worked in year 1 should also be present in year 2 *and* 30 percent of the employees in year 2 should have been employed in the establishment in year 1. In the latter, the definition of the same labor force is more restrictive.

Status variables provide detailed information on different establishment characteristics that are obtained during the November registration period. These status variables can be divided into (i) background related information and (ii) figures on the number of employees in the establishment. Background information gives insight in the type of

as founded in 1980.

industry in which the establishment is active, the geographical location of the establishment, the type of ownership, number of affiliated establishments, and the year in which the establishment is founded.

The year of founding can be identified from the first four digits in the establishment identification number. The industry in which an establishment is active is based on the four-digit NACE industry classification. Statistics Denmark has extended this classification by adding two digits, creating a more detailed six-digit DB93/DB03 classification. Within IDA there are two breaks in this industry classification. First, before 1992 there is no detailed industry classification for services. Second, in the year 2003, DB93 is updated to the DB03 industry classification; these changes are, however, minor.[8] The location of the establishment is identified by a municipality code from one of the 276 municipalities that existed in Denmark prior to the 2007 municipality and regional reform. Algorithms have been developed to group these municipalities in local labor market regions (Andersen, 2000, 2002); in addition, documentation is available on which municipalities are part of which larger administrative region (i.e. the former Danish counties or the in 2007 created regions). IDA also makes a distinction between a number of different ownership structures (i.e. sole proprietorship, partnership, limited partnership, government owned establishments, and foreign affiliations). Finally, for each establishment the number of affiliate establishments are identified (i.e. the total number of establishments that share the same firm identification number).

There are four variables that provide information on the number of employees in the establishment. First, the head count of employees active in the establishment. Second, the number of people for whom the occupation in the establishment is not the primary source of income. These two variables are based on the employees that worked for the establishment during the November registration period. The remaining two employment figures give information on the number of employees that have been active in the establishment during the last year. One variable identifying the number of employees measured in full time equivalent; the other presenting the head count of employees during the last year.

The last types of variables are the *retrospective* and *prospective variables.* These variables give information on the identity of the establishment in the previous and the following year. The retrospective variables identify whether the establishment was present in the database in the previous year or if the establishment is newly established. There are two ways in which an establishment remains the same compared to the previous year. Either the establishment does not experience any substantial change or the identity number remains the same but the establishment was involved in a split up or a merger. A new establishment appears in five different ways: (i) the establishment is newly founded, (ii) the establishment is separated from an already existing establishment, (iii) an establish-

[8] A key to translate DB03 to DB93 and vice versa is available on the website of Statistics Denmark. As of January 2009, a new DB07 industry classification is available with some more substantial changes.

ment enters due to the takeover of a building that belonged to an establishment that closed down and was active in the same industry, (iv) a self-employed individual that did not employ any employees in the previous year, and (v) an establishment that did not employ employees in the previous year. In the last two scenario's, the establishment already existed previously.

Table 1: IDA Establishment Variables

VARIABLE NAME	DESCRIPTION	YEARS
Identification Variables		
LBNR	Establishment identification number (The first four digits to indicate the year in which the establishment appeared for the first time in the November statement.)	1980-2006
ARBGNR	Firm identification number	1980-2006
Status Variables		
BRANCHE1	Industry classification DB93.	1980-2003
BRANCHE03	Industry classification DB03 (continuation of DB93).	2003-2006
ARBKOM	Code for the municipality where the firm is located.	1980-2006
EJERKO	Ownership structure of the establishment.	1980-2006
FILIAL	The number of establishments that are in the entire firm (max. 9).	1980-2006
ANTNOV	Employees in head count.	1980-2006
ANTNOVBI	Number of employees for who the occupation is not their primary source of income.	1980-2006
AARVRK	Employees in full time equivalent that during the last year have been employed in the establishment.	1980-2006
ANTAAR	Head count of employees that during the last year have been employed in the establishment	1980-2006
Retrospective and Prospective Variables		
IDTILB	Indicates the reason of a change in establishment identity number compared to the previous year.	1980-2006
IDFREM	Indicates the reason of a change in the establishment identity number compared to the following year.	1980-2001, 2003,2005
LBNRTILB	Indicates the related establishment identification number in the previous year.	1981-2006
LBNRFREM	Indicates the related establishment identification number in the following year.	1981-2001, 2003,2005
FRELTILB	Indicates whether the establishment is founded due to a firm internal separation or a separation where the establishment is active under an other firm identification number .	1980-2006
FRELFREM	Indicates whether an establishment is merged with an establishment within the same firm or acquired by an establishment in another firm.	1980-2001, 2003,2005

Note: the names of the variables might vary.

Another retrospective variable identifies the related establishment identification number in the previous year. This information is most valuable for those establishment that are founded as a results of a separation from an existing establishment. For these entrants it is possible to identify from which establishment it got separated; a newly founded firm will not be related to any previous establishment. Furthermore, the last retrospective variable indicates whether this separated establishment is still part of the same firm or whether it started to operate under a new firm identification number.

The prospective variables identify whether the establishment is still present in the following year or if the this establishment will disappear. The ways in which an establishment remains in the database is similar compared to the retrospective variable. Either the establishment does not experience any changes or the establishment will be involved in a separation or a merger but without losing too much of its identity for a change in identity number to occur. An establishment is not present in the database in the next year for the following five reasons: (i) the establishment closes down, (ii) the establishment is merged with or acquired by an already existing establishment, (iii) an establishment is closed down due to a transfer of their buildings to a new entering establishment, (iv) the establishment moves into a self-employed individuals without any employees, or (iv) an establishment moves into a situation in which it has no employees in the following year.

Two prospective variables remain. One variable indicates the related unique establishment number in the following year. This information is most valuable for establishments that will close down as a result of a merger because it is possible to identify the other establishment(s) in the merging process. Establishments that close down completely will have a related establishment identification number of zero. The last prospective variable indicates whether this is a merger of establishments within the same firm or whether the establishments were active under two separate firm identification numbers.

2.2 Employment level

The second data level provides information on the employment relationship of the Danish labor force. The employment information is, at least in the database used for this study, only available on the main occupation of the worker; even though one sideline occupation can be identified. The main occupation is that occupation which generates the employee's primary income. In a similar fashion as the establishment level, variables can be divided into three categories: identification variables, status variables, and retro- and prospective variables. An overview of the main variables available in this study is shown in Table 2.

In the employment database, three identification numbers are presented. Two of these identification numbers have been described earlier in Section 2.1 (i.e. the firm and the establishment identification number). By using these two numbers it is possible to link the employment level with the establishment level at any given year from 1980 and onwards. In some cases the establishment identity number is zero; this occurs whenever the employee is not physically assigned to an establishment (e.g. a person works from home or works at multiple establishments within one firm). Consequently, these individuals cannot be connected to a specific establishment. The third identification number is the personal identification number. This is a random number connected to a person's social

Table 2: IDA Employment Variables

VARIABLE NAME	DESCRIPTION	YEARS
Identification Variables		
PNR	Personal identification number.	1980-2006
LBNR	Establishment identification number.	1980-2006
ARBGNR	Firm identification number.	1980-2006
Status Variables		
ANSAAR	Year of employment in the current establishment.	1980-2006
TYPE	Variable indicating whether the individual is identified as employer or employee.	1980-2006
PJOB	Indicator whether the occupation is full or part time.	1980-2006
TILKNYT	Indication on the number of hours an employee works during the week.	1980-2006
TIMELON	Hourly wage of the employee.	1980-2006
TLONKVAL	The uncertainty factor in the hourly wage.	1980-2006
KMAFST	Distance between location of residence and location of the establishment.	1980-2003
Retrospective and Prospective Variables		
ANSXTILB	Indicates the employment status of the individual in the previous year.	1980-2006
ANSXFREM	Indicates the employment status of the individual in the following year.	1980-2003, 2005

Note: the names of the variables might vary.

security number and is fixed over time. For this reason, it is possible to follow the labor mobility patterns of individuals and identify their employment history.

The *status variables* provide information on the current employer-employee relationship. The first variable is the year in which a person entered the establishment. The second status variable makes a distinction whether the person is an employer or an employee. A similar, more detailed, variable is presented in the person level database that will be described in the following section. Two other variables describe whether the employment relationship is full or part time; the first variable only makes the distinction between full and part time while the other also lists the number of hours a person works during the week. Two other variables are linked to income, other income variables will be presented in the following section. One variable presents the hourly wage of the individual while another variables determines the uncertainty factor of this hourly wage. This uncertainty is negatively correlated with the number of hours a person works during the week. The last status variable indicates the distance in kilometers between the location of residence and the location of the establishment.

On the employment level there is one *retrospective variable* and one *prospective variable*. The retrospective variable identifies the employment status of the individual in the previous year. A person was either employed for the same establishment or entered the establishment from: an existing establishment in the same firm, from a different firm, from another establishment into a new founded establishment, from a situation of unemployment, from outside the labor force, from abroad, or from a period of leave. The prospective variable identifies the employment status of the individual in the following year. If a person is not present in the establishment in the following year the establishment, he or she will either, move to another establishment in the same firm, to another firm, move as a result of a closure of the establishment, move into unemployment, move outside the labor force (e.g. retirement), move abroad, take a period of leave, or die in the following year.

2.3 Person level

The last database in the vertical IDA structure is the one that contains information on the personal level. This person level database consists out of one person identification number, which can be used to merge this personal information with employment level information, and a various number of status variables. In this particular case, personal information is only available for those individuals that are part of the labor force. Therefore, IDA reports personal information on approximately 2.5 million individuals in any given year. Due to entry in and exit out of the labor market, personal information is available for more than four million different individuals over the entire period 1980-2006. An overview of the main variables used is shown in Table 3.

The *status variables* on the personal characteristics can be grouped in a number of different categories. One category includes the background variables from an ascribed nature. This category includes: gender, age, location of residence, citizenship, and country of origin. The location of residence follows the same municipality coding as the location of the establishment. Citizenship and country of origin are based on a list of over 250 different nation codes; furthermore, a variable is created that identifies whether the individual is a Dane, a first, or a second generation immigrant. It is even possible to identify family relationships between individuals; however, this database is not available in this study.

Another category of variables provides information on the education of the individual. The two most frequently used education variables are (i) the highest fulfilled education and (ii) the education the person is following at that point in time. These variables are coded with an eight-digit education code. The first two digits provide information on the level of education (e.g. high-school, bachelor, master, etc). The following two digits indicate the type of education (e.g. social science, humanities, engineering, etc). Digit five and six provide a more detailed grouping (e.g. economics and management, sociology, psychology, etc). The last two digits are the separate education titles available within each discipline. In addition to these two education variables, IDA provides information on the year the individual completed the education and the municipality in which the person is following an education.

A third category of variables are employment related variables. Within this category there is information on the current work relation (i.e. the industry in which the person works and the position the person holds within the firm). The variable that indicates the position the person holds is comparable with the variable in the employment level database although more detailed. Another group within the employment related variables provide information on the person work history (i.e. the total work experience and the degree of unemployment). On top of that, there is a category providing a number of income variables (i.e. taxable income, gross income, net wage, and surplus a person obtained from their firm).

Table 3: IDA Personal Variables

VARIABLE NAME	DESCRIPTION	YEARS
Identification Variables		
PNR	Personal identification number	1980-2006
Status Variables		
KON2	Gender	1980-2006
ALDER2	Age	1980-2006
BOPKOM	Municipality of residence.	1980-2006
STATKOD	Citizenship	1980-2004
IELAND	Country of origin.	1980-2002, 2004-2006
IETYPE	Indicates if a person is a Dane, first, or second generation immigrant.	1980-2002, 2004-2006
HFFSP	Highest fulfilled level of education.	1980-2006
IGFSP	The education the person is currently following.	1980-2006
HFAFGTP	The year the person finished his highest fulfilled level education.	1980-2006
UDDKOM	Indicates the municipality where the person is following the education.	1980-2003
PDB932	The industry in which the individual works.	1980-2003
PDB032	The industry in which the individual works.	2004-2006
PSTILL2	Type of position the person holds in the main occupation.	1980-2006
ASOCIO2	Socio-economic group	1980-1995
SOCIO	Socio-economic group (continuation of ASOCIO2)	1995-2002
SOCIO02	Socio-economic group (continuation of SOCIO)	2004-2006
ERHVER	Work experience from 1980 and onwards (in 1000s).	1980-2006
ERHVER79	Work experience before 1980.	1980-2006
ARLEDGR	Degree of unemployment throughout the year.	1980-2006
LEDPERI	Period of unemployment.	1980-2006
SUMGRAD	Sum of unemployment (from 1980).	1980-2006
BRINDK2	Gross income.	1980-2006
LONIND	Net income.	1980-2006
SKPLIND2	Taxable income.	1980-2006
OVSKVI	Surplus from the firm.	1980-2004
OVSKVIR	Surplus from the firm. (Continuation of OVSKVI)	2005-2006

Note: the names of the variables might vary.

3 Databases to be Merged with IDA

Another strength of IDA, besides its universal and longitudinal characteristic, is the possibility to merge the database with a wide range of existing databases. This can be databases that are available at Statistics Denmark, surveys made in collaboration with Statistics Denmark, or databases available at other organizations. A requirement of the latter is the presence of a social security or a firm registration number because these databases can only be merged based on these identification numbers. The most remarkable mergers I came across are a merge between IDA and the medical database (Dahl, 2009), a merge between IDA and a survey among Danish entrepreneurs and wage earners in an attempt to identify the characteristics of Danish entrepreneurs (Dahl et al., 2009), and a merge between IDA and a database of Danish merchant marines (Isakson, 2009). In this section, I will illustrate this merging process by describing the merge with two type of databases that I have used in the past. The first database contains firm level accounting data. The second database is an innovation survey among Danish firms in 2006 focusing on organizational and technological change.

3.1 Firm Level Accounting Data

The database on firm level accounting data is split up in two types covering two separate periods. One database that covers the period 1992 to 1999 and a second version that covers the period 1999 and onwards. These databases contain a various number variables: number of employees (FTE and head count), industry classification (DB93 or DB03), municipality code, equity, fixed assets, purchases, turnover, profits, exports, paid wages, and value added.

There are, however, some limitations connected to this database. First, this database has a triviality limit, which means that this information is not available for all firms within Denmark, despite this loss of observations, many firms remain in the database. Second, information is only available on the firm level; as a result, the performance of individual establishments in multi-establishment firms cannot be determined. Nevertheless, approaches to deal with this problem have been developed. One example is to use the distribution of wages between the different establishments to account for the distribution of other financial variables, (e.g. value added) (Boschma et al., 2009). A third limitation of the database is the break in the data. The most visible break is between the two different databases, i.e. 1999; there also appears to be a break in 1995 when the total number of firms suddenly increases. As a result of these breaks, one can only use these firm statistics for shorter periods of time because a comparison between the periods that are divided by a break are hard to make. The break in the data will not be a problem in the analyses that follow because I only use the general firm statistics from 1999 and onwards.

This databases can be merged together with IDA based on a firm identification number.

This firm identification number is slightly different compared to the firm identification number (i.e. ARBGNR), which I presented earlier; nevertheless, Statistics Denmark provides a key that creates a link between these two firm identification numbers.

3.2 Danish Innovation Survey

DISKO4 is the most recent DISKO survey on organizations, employees, and research and development strategies in Danish firms survey conducted in 2006 by Statistics Denmark on behalf of four research groups (IKE, CARMA, CIP and CCWS) at Aalborg University.[9] Just like the previous DISKO surveys, DISKO4 takes its point of departure in firms that participated in earlier DISKO survey augmented with firms in those industries that can be found in general firm statistics of 2004 (Statistics Denmark, 2007).

This survey is a follow up of previous surveys in the so-called DISKO project, which have been conducted in 1996 (DISKO1), 2001 (DISKO2), and 2004 (DISKO3), that focuses on organizational and technological change in Danish firms; Reichstein and Vinding (2002) presents a description on the earlier versions of DISKO. The questionnaire was sent to a total of 4,136 Danish firm that were selected on a number of criteria. The first criterion was to include those firms that participated in previous DISKO surveys. In total 1,552 firms were identified as still being operational. The second criterion was to include all the firms with more than 100 employees and finally a unbiased selection of firms in the size category 20-49 and 50-99 employees (Statistics Denmark, 2007). Table 4 provides an overview of the distribution of these 4,136 firms based on size and industry.

Although the DISKO surveys have a general requirement to include firms larger than 20 employees there are 229 firm that are smaller. These are firms that participated in the previous DISKO surveys and during the years became smaller. Despite this fact a decision has been made to include them in the large sample.

Statistics Denmark received 1,781 questionnaires from these 4,136 firms 1,781, 6 of these questionnaire turned out to have been answered already so in total there are 1.775 unique questionnaires resulting in a response rate of 42,9 percent. Eventually, 1770 observations were included in the database. The distribution of the answers, based on the same categorization as in Table 4, are presented in Table 5.

Just as any merging process, the merge between DISKO4 and IDA is not a straightforward exercise. In this particular case there are two main challenges. First, The merge between IDA and DISKO4 is based on the same firm identification number as described in Section 3.1; consequently, in the case of a multi-establishment firm, there is a need to determine which establishment has filled out the questionnaire. This survey has been sent to the head office of each firm. The problem is that the database does not point out

[9] In Timmermans (2008) an English version of the DISKO4 questionnaire is available.

Table 4: Distribution of the DISKO4 Sample based on Industry and Size

INDUSTRY	SIZE GROUPS				TOTAL
	LESS THAN 20 EMPLOYEES	20-49 EMPLOYEES	50-99 EMPLOYEES	OVER 100 EMPLOYEES	
Manufacturing	37	342	415	576	1370
Construction	58	217	125	78	478
Trade and Repair	79	401	294	266	1040
Hotel and Restaurants	6	45	29	23	103
Transport	24	107	97	124	352
Financial Services	1	21	33	80	135
Business Services	24	197	161	201	583
Culture and Sports	0	24	32	19	75
Total	229	1354	1186	1367	4136

Source: Statistics Denmark (2007); Timmermans (2008)

Table 5: Distribution of the DISKO4 Questionnaire based on Industry and Size

INDUSTRY	SIZE GROUPS				TOTAL
	LESS THAN 20 EMPLOYEES	20-49 EMPLOYEES	50-99 EMPLOYEES	OVER 100 EMPLOYEES	
Manufacturing	14	190	177	210	591
Construction	26	97	53	38	214
Trade and Repair	34	169	116	98	417
Hotel and Restaurants	2	16	14	5	37
Transport	12	45	38	48	143
Financial Services	0	9	17	43	69
Business Services	10	91	71	97	269
Culture and Sports	0	12	10	8	30
Total	98	629	496	547	1779

Source: Timmermans (2008)

which establishments are regarded as head offices. In this particular case, the largest establishment is identified as the head office.

Whenever there is the desire to merge a survey with IDA there is the challenge of timing. It takes approximately two years before any given year is made available in the IDA database. In this particular case, the information on the year in which the survey was conducted, i.e. 2006, was only made available at the end of 2008. To correctly determine the performance and composition of the firm requires some patience on those questions that apply for the year in which the survey was conducted. After dealing with these, and other minor merging issues, DISKO4 is supplemented with a vast array of background variables.

4 Limitations

Despite the vast number of strength and opportunities that IDA provides for researchers there are a number of limitations that needs to be considered and that have been sparsely mentioned throughout this document. One limitation is the analysis over a longer period of time. As I already mentioned, there are several breaks in several variables that are available in the database. Consequently, it becomes difficult to do an analyses on the entire time span of the database.

Furthermore, there is the fact that there are only yearly observations for each establishment. Due to these yearly snapshots, it is not possible to identify what occurs between the different November registration periods. During this period, short-term period of employment (e.g. between January and July) cannot be observed, which is especially interesting for those industries that rely on season workers. In addition to employment the analyses of firm dynamics (i.e. founding, growth and disbanding of firms) can only be determined on a yearly basis. Firms that founded and disbanded in between the different November registration periods are not observed in the database. Studies on entrepreneurship will lose many short term entrepreneurship activities.

Another shortcoming of the database is the reliance on register data. Despite that much information can be obtained from this approach, it misses those variables that indicate the modes of interaction within firms (e.g. modes of work, interaction between employees, strategic decision making processes) and how the firm interacts with the environment (e.g. collaboration with customers and suppliers, involving third party consultants, receiving government support, interaction with competitors, etc.), which are very crucial in determining the performance of firms. This variables can provide a picture on the daily operations of the establishments. Although not present in the database, many of these variables can be obtained by combining this database with more qualitative approaches.

A strength and main purpose of the database was to identify the mobility of workers; however, the motivation for this mobility cannot be determined. A move can be initiated by the employer (e.g. lay off) or the employee (e.g. resignation), although a potential motive can be derived by looking at some other variables (e.g. wage level, firm performance, personal characteristics, etc). In addition, the recruitment of employees into the organization is another factor that cannot be determined. The motives for leaving a firm and joining another can be regarded as an important indicator for future firm performance.

Another concern is related to the activity of the firm. The only indicators that hint upon the activities of the firm are the industry classification codes. However, the exact nature of their activities (e.g. the products they produce, services they offer, and whether these activities are characterized as intermediate or final goods or services) is unknown.

Furthermore, Christensen (2008) has shown that the assigned industry classification often does not represent the true activities of a firm. For this reason, it is hard to identify the firm's real competitive environment. At least the competitive environment on the output side. From an input perspective some competitive pressures can be identified (e.g. the competitive pressures on the labor market) Sørensen (2004).

5 Concluding Remarks

The purpose of this document was to provide a description on the structure and content of the Danish Integrated Database for Labor Market Research. I have not, despite the fact that the database is used in a vast array of scientific disciplines, encountered an English written document on the database, which was the motivation for writing this document. For the description of this database, I relied on the information made available by Emerek et al. (1991) and the documentation, in Danish, on the homepage of Statistics Denmark.

Up to know, the database has for me been a source of inspiration for various type of research, both on the level of the firm as on the regional and national level. It is thus not surprising that there is a broad international interest in using this and similar databases. The explanatory power of these databases increases for several reasons: (i) more years are added, (ii) more countries grant access to IDA-like databases creating the opportunity to do comparative work, (iii) dedicated surveys are being developed, which combine this universal and longitudinal data with in-depth qualitative data, and (iv) more sophisticated methods are applied to study those phenomena that can only be studied by using such sources of data.

References

Andersen, A. K. (2000). Commuting areas in denmark. *AFK Forlaget.*

Andersen, A. K. (2002). Are commuting areas relevant for the delimitation of administrative regions in denmark. *Regional Studies*, 36(8):833–844.

Boschma, R., Eriksson, R., and Lindgren, U. (2009). How does labour mobility affect the performance of plants? the importance of relatedness and geographical proximity. *Journal of Economic Geography*, 9(2):169–190.

Campbell, B. A. (2005). Using linked employer-employee data to study entrepreneurship. In Alvarez, S. A., Agarwal, R., and Sorenson, O., editors, *Handbook of Entrepreneurship Research*, volume 2, pages 143–166. Springer.

Christensen, J. L. (2008). Questioning the precision of statistical classifications of industries. *DRUID Summer Conference 2008.*

Dahl, M. S. (2009). The cancer of organizational change. Paper presented at the DRUID Summer Conference 2009.

Dahl, M. S., Jensen, P. G., and Nielsen, K. (2009). *Jagten på fremtidens nye vækstvirksomheder*. DJØF Forlag.

Emerek, R., Vejrup-Hansen, P., and Leth-Sørensen, S. (1991). Ida- en integreret database for arbejdsmarketforskning. (In Danish).

Eriksson, T. and Kuhn, J. M. (2006). Firm spin-offs in denmark 1981-2000: patterns of entry and exit. *International Journal of Industrial Organization*, 24:1021–1040.

Isakson, C. (2009). Entrepreneurs in sticky clusters: Location choice of a nomadic labor force. In *Paper presented at the DIME DRUID Academy Winter 2009 PhD Conference*.

Reichstein, T. and Vinding, A. L. (2002). Documentation of the ida-disko database. *Department of Business Studies - IKE Group, Aalborg University*.

Sørensen, J. B. (2004). Recruitment-base competition between industries: A community ecology. *Industrial and Corporate Change*, 13(1):149–170.

Statistics Denmark (2007). Metoderapport til disko 4-undersøgelsen. (In Danish).

Timmermans, B. (2008). Documentation on the disko4-ida merge and the creation of the panel dataset disko2-disko4. *I3 Working Paper Series*, (24).