

Microdata for research and analysis – potential and problems

Flemming Petersson, Statistics Denmark (flp@dst.dk)

Jørn Korsbø Petersen, Statistics Denmark (jkk@dst.dk)

Ole Schnor, Statistics Denmark (osc@dst.dk)

Leif Husted, Institute of Local Government Studies, Denmark (lh@akf.dk)

Paper for the Siena Group on Social Statistics, meeting in Helsinki February 9-11, 2005 Session C: Record Linking presented by Flemming Petersson, Statistics Denmark

1. Data potential and accommodation of external researchers

The vision of Statistics Denmark is to be among the leading in the world in terms of utilisation of microdata for research purposes, and Danish researchers now have a unique opportunity to include the microdata of Statistics Denmark in their research. There are several reasons why this vision seems to have been realised here in 2004.

Remote access is a great advantage

In December 2002, the Board of Statistics Denmark decided to grant researchers 'remote access' to analysis of microdata. Until then, an on-site scheme was available where the researchers had to be physically present at the premises of Statistics Denmark to use our microdata. The remote access scheme means that researchers can now remain in their own research environment. Obviously, this offers substantial time savings for the individual researcher. Working with the microdata of Statistics Denmark has thus become a natural element of the daily work of many researchers. Access to the system is subject to authorisation of the research institution and approval of the individual research project. For details on authorisation rules and security matters concerning remote access, please refer to *From on-site to remote data access – the revolution of the Danish system for access to microdata*. (Otto Andersen)

General grant 2002-2005

A general grant for the period 2002-2005 from the Ministry of Science, Technology and Innovation have made it possible to provide researchers with low-price data extracts and free computer runs. This means that a large number of research projects with limited budgets, such as PhD projects, can also be realised. A specialised Division for Research Service has been set up at Statistics Denmark with the task of providing the best possible service to researchers. The Division comprises 12 persons, two of whom work at our Århus branch.

Growing number of researchers and projects

On that basis, the number of researchers and projects has increased. In 2002, there were 177 active researchers in all, and this number increased to 235 in 2003. Similarly, the number of active projects in 2003 was 170 compared to 114 in 2002. This growth continued in 2004, and the second quarter saw the highest activity ever, cf. table 1.1.

Table 1.1. Quarterly distribution of active projects and researchers

Year	Quarter	No. of active projects	No. of active researchers
2002	1	60	91
	2	66	108
	3	79	115
	4	85	127
2003	1	107	153
	2	106	136
	3	107	144
	4	122	164
2004	1	120	159
	2	131	173

Many steps have thus been taken in recent years to create the best possible technical and financial framework for researchers' work with microdata.

Registers are the foundation

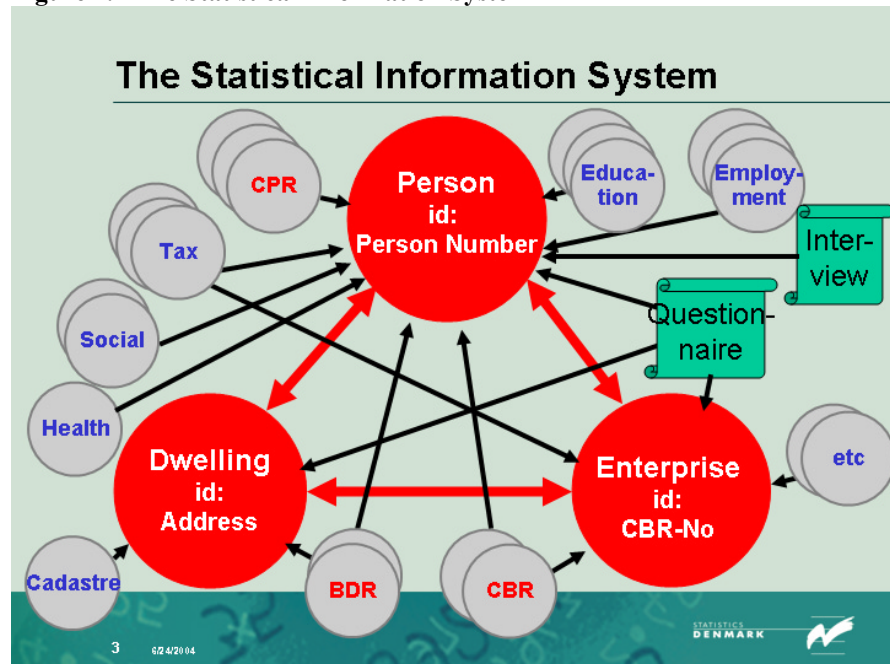
The primary reason for the great success of the research scheme, however, is the comprehensive collection of statistical registers at Statistics Denmark. Without this foundation, the system would not exist.

The statistical registers build on administrative registers and have the following characteristics:

- they contain high-quality data
- they often comprise the entire population
- they cover several years
- they can be linked through a set of keys

In the context of research, the collection of registers constitutes a ‘goldmine’ of data, allowing longitudinal analysis of populations along with a large amount of background information.

Figure 1.2 The Statistical Information System



Great potential

As shown in figure 1.2, the registers are linked by various keys for individuals, families, dwellings, workplaces and enterprises. The possibilities of combining the data are nearly indefinite, which is illustrated by the many diverse research projects carried out in the fields of economy, sociology, demography, epidemiology, etc. In addition to the statistical registers, the Division for Research Service has the Register of Medicinal Products at its disposal, which offers unique research possibilities in the field of medical science. In projects based on the researchers’ own surveys, the researchers have the option to include background information from the ‘goldmine’ of data.

Anonymised keys

In respect of all projects, the Division for Research Service handles the linking of the registers and any survey data. The Division ensures that any keys and variables that can identify the individual observation are either anonymised or deleted before the data are made available to the researcher. Plain text with names, addresses and the like are deleted, and person numbers, CBR numbers and address codes are anonymised beyond recognition.

Research reveals weaknesses and problems

The data potential is large, but the work of the Division for Research Service and researchers also reveals some of the problems that occur when linking the many statistical registers. In this paper, we will discuss some of these problems. Section 2 deals with social statistics and section 3 concerns business statistics. Section 4 points to some of the problems that arise in connection with longitudinal analysis,

section 5 emphasises the importance of consistent documentation, while section 6 discusses the possibilities of analysing behaviour in continuous time.

2. Individuals and families

Projects in social statistics

The vast majority of the research projects apply data from the field of social statistics, and the majority of projects relate to social and health science.

The projects are usually based on a delimited population, e.g., all residents in nursing homes. The population is typically identified on the basis of the person registers at Statistics Denmark, but sometimes the population derives from a research institution, for instance in the form of a survey. The population is then linked with register data, such as information on families, incomes, socio-economic status, education, etc.

Example

One example from the field of social science is the project called *Hjemløses vej ud af hjemløshed* ('the way of the homeless out of homelessness'). The project objective is to examine what types of efforts have improved the situation of the homeless. It is therefore important to have information on the homeless both before and after their becoming homeless. The population of homeless people derives from survey data collected by the research institution. The homeless are to be observed over a number of years in terms of family conditions and a wide range of background information such as housing conditions, labour market ties, crime, incomes and transfers, social benefits to children and young people and medical information from 1980 and onwards. The wide range of information is decisive to creating an overall picture of their situation both before and after they became homeless.

Person number

The researchers' need to be able to observe people over a longer term places special demands on the identification of the individual – the person number, which is recorded in the Civil Registration System. Unambiguous identification of each person thus requires that the person number does not change over time. Normally we assume that the person number is an unambiguous and stable key that follows a person through his or her entire life, but as shown below, some persons change their person number up to several times during their lives. In connection with research projects, this may give rise to special problems in relation to observing persons over time.

Change of person number and research projects – examples and scope

The mortality and occupational survey of Statistics Denmark illustrates the problem of non-unambiguous identification of persons, i.e., a change of person number.

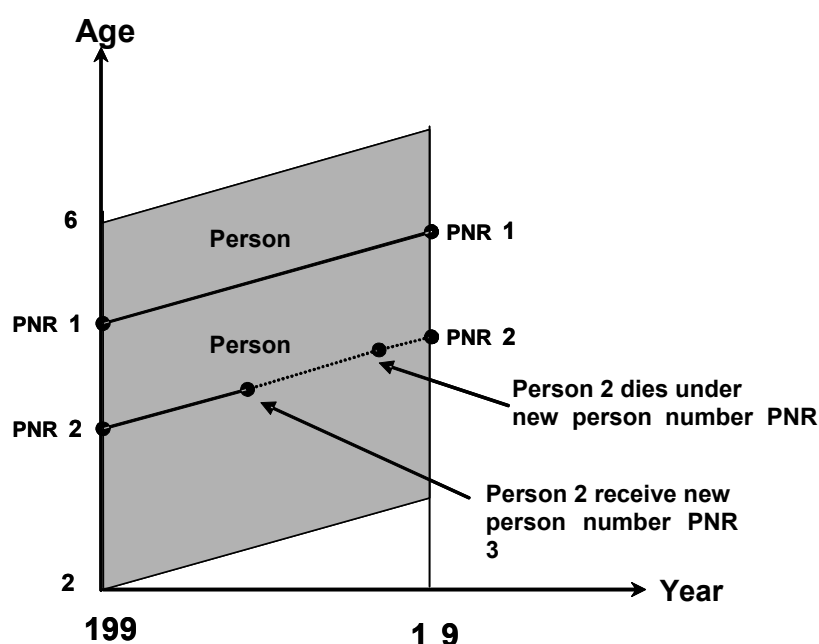
In the mortality and occupational survey a cohort of persons is observed for five years, e.g., from 1 January 1991 to 31 December 1995, to measure the differences in mortality between different occupations.

The population is extracted from the Population Statistics Register and comprises persons aged 20 to 64 years at 1 January 1991. These persons are then observed during the five-year period in terms of emigration and deaths. This means that only the person number assigned to the individual person *at the beginning of the period* is included in the survey.

Correct recording of the events requires that the persons included in the survey do not change their person number. If a person changes his or her person number during the survey period and subsequently dies or emigrates, the event will be re-

corded under the *new* person number. Accordingly, events that are essential for the survey will not be recorded. The figure below illustrates the problem.

Figure 2.1 Illustration of change of person number



The figure follows two persons illustrated by a 'time-line' each. Person 1 has the person number PNR 1 in the survey, while person 2 has the person number PNR 2 at the beginning of the period.

Person 1 remains in the population for the entire period, meaning that this person neither dies nor emigrates. Also, person 1 has the same person number (PNR 1) for the entire period.

Contrary to this, person 2 changes his person number during the period from PNR 2 to PNR 3, and he subsequently dies during the survey period. The death is recorded under the new person number, PNR 3, which is not known in the survey. The observation of death is therefore lost, and person 2 (who is only identified by PNR 2 in the survey) will – erroneously - appear to have remained in the population for the entire period like person 1.

The problem of a change of person number is of a general nature when a specified population is to be observed over a longer term.

If we look at the population residing in Denmark in the period 1980 to 2002, 1.27 per cent changed their person number once since 1968, cf. the table below.

Table 2.2. Change of person number since 1968 for persons residing in Denmark from 1980 to 2002

No. of person number changes	Persons	Per cent
0	7,187,615	98.65
1	92,876	1.27
2	4,875	0.07
3	400	0.01
4+	100	0.00

Note: The table comprises all persons who have changed their person number since the establishment of the Danish Civil Registration System (CPR) in 1968.

Who changes person number?

As at 1 January 2003, 56,738 persons of the 98,251 persons who changed their person number still reside in Denmark.

In the table below, these persons are divided into the group 'Danish', which includes persons with at least one parent born in Denmark, and the group 'Immigrants/descendants'. For the purpose of comparison, the entire population is grouped accordingly.

Table 2.3. Distribution into the groups Danish and Immigrants/descendants of persons with person number changes and the entire population residing in Denmark at 1 January 2003

Danish-immigrants/descendants	Persons who changed their person number	Per cent	Entire population	Per cent
Danish	46,586	82.1	4,952,818	92.0
Immigrants / descendants	10,152	17.9	430,689	8.0

The table shows that a relatively large share of immigrants/descendants change their person number in relation to their aggregate share of the population. The group of immigrants/descendants thus accounts for 18 per cent of the person number changes, while together they only represent 8 per cent of the population.

In the table below, the 10,152 immigrants/descendants with a change of person number are distributed by country of origin. The table illustrates a higher incidence of person number changes among persons from less developed countries in relation to their aggregate share of the group of immigrants/descendants.

Table 2.4. Immigrants/descendants with person number changes by country of origin

Countries of origin	Immigrants/descendants who changed their person number	Per cent	All immigrants/descendants	Per cent
More developed countries	2,816	27.7	182,001	42.3
Less developed countries	7,223	71.2	246,412	57.2
Not stated	113	1.1	2,276	0.5

Why do people change their person number?

There are several reasons for a change of person number, such as wrong recording and lack of documentation. Persons who have no identification papers at their entry into Denmark, for example, may be recorded under the wrong date of birth. Another common mistake is faulty recording of the sex at birth by the hospital. Moreover, a person who has a sex change operation is always assigned a new person number.

Person number changes often take place in connection with contact with the local authorities, for instance because a person is applying for pension payments and therefore has to provide evidence of his or her date of birth. Occasionally, the correct person number is not assigned to the person until his or her death.

How to solve the problem?

The problem of person number changes has been solved in the Demographic Database, which contains detailed information on the demographic events of the population such as change of address, migration, change of marital status, etc., from

1980 and onwards. Since each person is observed at a very detailed level, unique identification of all persons over time is crucial. Instead of using the person number, it has been decided to use a serial number, which is unique for each individual person and does not change over time. No matter how many times a person changes his or her person number, he or she will always have the same serial number.

3. Employees, workplaces and enterprises

Great demand for business data

The number of projects that want to include business statistical data in their analyses is growing, and the requests for data are often somewhat beyond the scope of the present business statistics. The requests typically relate to financial key figures for the enterprise (e.g., turnover, value added and investments), information on workplaces (e.g., industry and number of employees), information on the employees (e.g., pay, work experience and education) information on ownership (e.g., member of a group), and the possibility to observe enterprises, workplaces and employees over time.

Information is requested at the following levels:

1. Employees (with reference to workplace)
2. Workplaces (with reference to enterprise)
3. Enterprises (with reference to group, if any)
4. Groups (with reference to foreign companies, if any)

Research includes Danish entrepreneurs

One example is the Danish research institution, *Centre for Economic and Business Research (CEBR)*, which analyses Danish entrepreneurship in its research project entitled *Entrepreneurship, Human Capital, and the Labour Market*:

"The entrepreneur is generally believed to play an important role in the modern economy as an engine of growth and as a creator of innovations. Although detailed descriptive studies of entrepreneurs and their activities are already in place – at least in the case of Denmark – our understanding of the involved issues is still far from perfect. This research project therefore focuses on the causes and consequences of entrepreneurship in Denmark."

For the project, data are retrieved from a number of databases at Statistics Denmark, the Integrated Database for Labour Market Research (IDA), Accounting Statistics, Enterprise Statistics and the Entrepreneurs Database. Personal data on the entrepreneur, such as family background, educational background and work experience, are key elements of the analysis, but information on the entrepreneur's firm and workplaces is also essential to get an overall picture. This may include the geographic location and industry of the workplaces, the financial position of the enterprise and not least the survival of the enterprise in the years following its establishment. For the project, data are therefore extracted from levels 1 to 3.

No information on groups

No systematic data exist at group level that can be made available to researchers. Much attention is being paid to this area at EU level, however, so in the long term researchers will be able to include the group level in their analyses. In the following, we will examine the data potential for levels 1 to 3.

IDA keeps track of employees and workplaces

The Integrated Database for Labour Market Research (IDA) covers the period 1980-2002 and contains a large number of person, family and workplace variables, amounting to a total of about 300 variables. The database was developed for re-

search purposes, and nearly all research projects include IDA variables to some extent. Many of the person variables in IDA are retrieved from other person registers at Statistics Denmark, but IDA is the only place where workplaces and employed persons can be observed from 1980 and onwards, and it offers a great potential for Danish labour market research. For the CEBR project, information was supplied about IDA's workplaces and employed persons for the whole period and all years, i.e. from 1980 to 2001.

Employment at each workplace is recorded annually at the end of November. This correlation is available in the Register-based Labour Market Statistics (RAS), while workplace identity is formed in IDA itself. Changes in workplace identity are described both in relation to the preceding year (maintained, created or separated) and the following year (maintained, discontinued or absorbed). A set of rules has been adopted for determining when a workplace changes identity, and at this point information is included on ownership, industry, labour force and address.

FIDA is the key to the enterprise level

Linking workplaces to enterprises is performed by means of the 'Enterprise/IDA key', known as FIDA. The quality and temporal scope of FIDA, however, is subject to the limitations presented by the statistics at enterprise level. Consequently, FIDA only covers the period 1995-2001. Before 1995 it is not possible to link enterprises to workplaces. To achieve the correct key, comprehensive data editing is required, including manual checks of the largest enterprises. Data editing is necessary because of the complex ownership structure of several major groups with many legal and administrative entities. The employment data in IDA may, for instance, be recorded under one entity, while the Enterprise Statistics may record the information under a different entity of the same group. Checking all entities in the two databases would be too time-consuming and therefore FIDA is not an absolutely correct key. Research projects have to take this uncertainty into account, e.g., by omitting the enterprises where discrepancies exist between IDA employment and turnover in the Enterprise Statistics.

Major changes in statistics at enterprise level

The social statistics are based on registers, the correlation between which has been largely unchanged since the early 1980s. Contrary to this, statistics at enterprise level have been subjected to many changes throughout the 1990s.

Enterprises (as legal entities) were not recorded under an unambiguous number until the establishment of the Central Business Register (CBR) in 1999. Before 1999, several administrative numbering systems existed, which complicated any integration with the Business Statistics.

The statistical methods of the Accounting Statistics, which constitute the central financial statistics for enterprises, have changed, and the statistics have not offered full coverage in terms of industry. The quality of the statistical method was improved substantially in 1995, but not until 1999 did the improved statistics cover all major industries (except agriculture and a few service industries). The good coverage of industries in the Accounting Statistics paved the way for a much needed quality improvement of the Enterprise Statistics as of 1999. For the period 1993-1999, the accounting data of the old Enterprise Statistics are "synthetic" at micro level, meaning that the data on the individual enterprise were calculated on the basis of the turnover of the enterprise, and accounting relations for strata were formed on the basis of the industry and number of employed persons. Moreover, the old Enterprise Statistics for 1995-1999 only cover industries which are subject to VAT. As of 1999, the new Enterprise Statistics cover all industries.

Time series breaks cause problems for researchers

The existence of more than one version of both accounting and enterprise statistics limits the possibility of obtaining consistent business data over a longer period. In a research context it is often a major problem that solid data are available only for

short periods. For the CEBR project the data supplied from the old Enterprise Statistics only cover the period 1995-1999.

Lack of demographic data at enterprise level

Furthermore, no method has been developed to determine the identity of an enterprise over time. In the business statistics mentioned above, the entity is the legal entity, which means that the 'birth' and 'death' of an enterprise depend entirely on the administrative recording of CBR number in the Central Business Register. If, for instance, an enterprise changes ownership/owner, then a new CBR number is created, and the enterprise will appear as a discontinued entity as well as a new entity in the statistics. The Business Statistics division is currently working on the methods for business demographics, and in time this will create whole new options for researchers to observe enterprises over time.

New enterprises recorded since 1990

The Entrepreneurs Database has recorded the number of actual, new enterprises and their 'survival' since 1990, which is beneficial to researchers. The annual number of new enterprises ranges between 14,000 and 18,000. These statistics enabled researchers to delimit the entrepreneurs for the CEBR project.

4. Demand for data for longitudinal analysis

Longitudinal data at individual level

The issues dealt with by researchers often give rise to special demands for data. Particularly so since research in numerous fields seeks to analyse the behaviour of individuals, which often requires access to data that observe individuals over time. Researchers' demand for data is thus highly focused on access to individualised data for a long, continuous period of time. Researchers therefore very much depend on the continuity of data. For this reason, breaks in time series constitute a major problem in a research perspective. New and improved statistical methods are certainly welcome, but if the consequence is a loss of temporal comparability, researchers will be very concerned about such initiatives. Naturally, researchers are also interested in improving the quality of data, but particularly so if Statistics Denmark also includes historical figures at the revision of its statistics so as to avoid any time series breaks. This was the case when the education statistics were revised a couple of years ago.

5. Documentation

Documentation facilitates the work and prevents misinterpretation

An important aspect for researchers is the existence of detailed documentation on the register data applied in their research. One of the reasons is that good documentation facilitates research, but the primary reason is that good documentation contributes to improving the quality of research, as it reduces the risk of misinterpretation. Nobody wants to publish results or propose policies that are based on useless or misinterpreted data.

Documentation must be available also for historical data

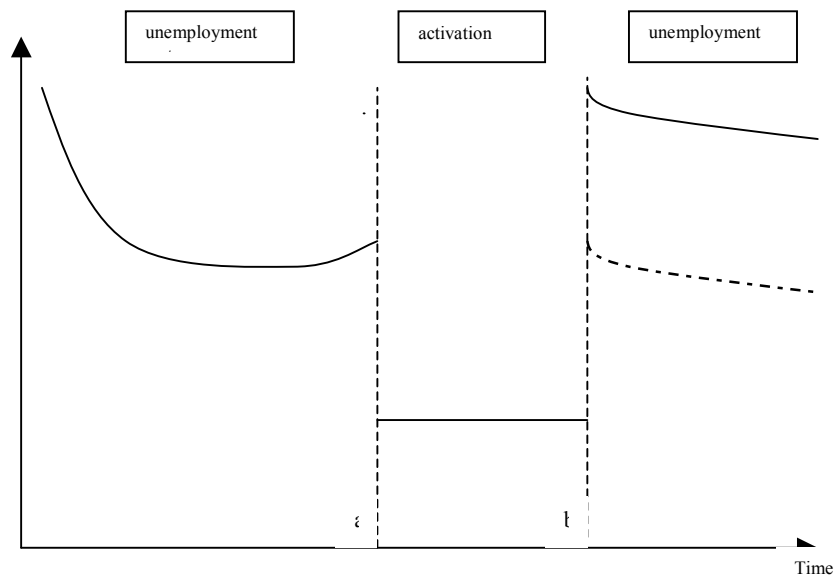
Today, documentation exists on the register data at Statistics Denmark, but the documentation could still be improved from the perspective of research. Firstly, documentation is needed on all registers applied by researchers, or at least on those applied most frequently. Secondly, the documentation should be consistent so that the names of variables in the data sets provided to researchers correspond to the names used in the documentation. Large parts of the documentation are already available at the website of Statistics Denmark, which is updated and expanded regularly. The largest problem, however, is the fact that the existing documentation focuses on the most recent data, while the historical documentation of data is inadequate.

6. Analysis of behaviour in continuous time

Focus on dynamic processes

In the field of social science, the past 15 to 20 years have seen a strong focus on dynamic processes such as studies of the movements of individuals between employment, unemployment and activation. Such analyses place special demands on data, as the timing of events is crucial; similarly, it is important to have exact information on the duration of each condition for the individual. For example, it is now common to analyse the duration of periods of unemployment and the aspects determining whether individuals are unemployed for shorter or longer periods of time. Other examples are analyses of the integration of immigrants, focusing on the duration of time from the immigrants' entry into Denmark until they get a job, start an education or no longer receive government benefits, and analyses of the effects of activation measures. The latter is a particularly good example of the importance of knowing the exact timing of events. Figure 6.1 illustrates how the probability of getting a job may be affected partly by the duration of unemployment, partly by the individual participating in a given activation measure.

Figure 6.1



The figure shows the probability of an unemployed person getting a job at different points in time during a period of unemployment, the horizontal axis showing the time and the vertical axis showing the probability of getting a job at a given time. At the beginning of the unemployment period, the probability of getting a job is high, but it decreases fairly rapidly. The probability is then nearly constant until shortly before the time a . At this point, the probability increases slightly again, as the person knows that, at the time a , he is to participate in activation and is therefore extra motivated to look for a job on his own. This 'motivation effect' was recently demonstrated by analyses based on Danish register data. During the period from a to b , the person participates in an activation measure, and the probability of getting a job in this period is very low. After the time b , the probability increases again. The increase in probability depends on the degree to which the person benefited from the activation measure. The upper line illustrates the effect of having participated in an effective measure, while the lower, dashed line shows the effect of having participated in a less effective measure.

To make an empirical analysis of whether a motivation effect actually exists and to analyse which measures are the most effective, the data required should indicate,

within small time intervals, the condition in which the person is, and for how long the person has been in that particular condition. In practice, the information is stated on either a weekly or monthly basis. To make an analysis as described above, the researchers will, as a minimum, require information covering a number of years on unemployment, employment, participation in activation measures distributed by type, for each week/month of the period.

Researchers developing suited data themselves

It is not possible to acquire directly applicable data for analyses such as the ones described above, since Statistics Denmark does not possess such data. It is possible, however, for researchers to buy the basic data at Statistics Denmark for constructing such data themselves. But this is no simple matter. Independently of each other, different institutions have spent considerable resources on constructing such longitudinal sequences where a person is observed on a weekly or monthly basis. Most of the institutions base their algorithms on various information from Statistics Denmark, but the algorithms of the individual institutions and the underlying, ad hoc definitions are made independently. No comparison has been made of the algorithms of the different institutions, so there is no overall knowledge of this field. This is, of course, a waste of resources, as lots of duplicate work is performed, and indeed the result of a joint, overall effort would most likely be better. Another unfortunate effect is the fact that the analyses made by different institutions are not directly comparable because they are based on different data.

Prioritising information

As mentioned, different research environments construct their own longitudinal sequences at the level of weekly or monthly data. The primary data sources applied for this work is information on unemployment from either CRAM (Central Register of Labour Market Statistics) or SSH (Coherent Social Statistics), information on pension, leave, maternity, sickness benefits and cash benefits from SSH, information on activation measures from AMFORA (statistics on labour market policy measures), education information from BUE (education and employment register), and employment information from either IDA or CON (salary information register). It is evident that, when combining information from so many sources, information about several different activities will often occur in a given week/month. This may be acceptable and also manageable in some cases. But often it will be necessary to prioritise the pieces of information by their importance. In other cases, there are conflicting data, and in those cases it is necessary to choose which pieces of information are the most reliable.

Table 6.1

<i>Month</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>	<i>13</i>
Unemployment	0	0	0	L	L	L	L	0	0	0	0	0	0
Pension	0	0	0	0	0	0	0	0	0	0	0	0	0
Leave	0	0	0	0	0	0	0	0	0	0	0	0	0
Maternity	0	0	0	0	0	0	0	0	0	0	0	0	0
Activation type 1	0	0	0	0	0	A1	A1	A1	A1	A1	A1	0	0
Activation type 2	0	0	0	0	0	0	0	0	0	0	0	0	0
Education	0	0	0	0	0	0	0	0	0	0	0	0	0
Employment	B	B	B	B	B	0	0	0	0	B	B	B	B
Condition	B	B	B	L	L	L	L	A1	A1	A1	A1	B	B

Table 6.1 illustrates an example of prioritisation of different pieces of information, the information at the top of table taking priority over information further down in the table, i.e., unemployment takes priority over pension, pension takes priority over leave and so on. At the bottom of the table, we see the condition arrived at by the algorithm for each month.