Statistical Journal of the IAOS 31 (2015) 249–256
DOI 10.3233/SJI-150894
IOS Press

Micro data integration for Labour Market Account

Pernille Stender*, Thomas Thorsen and Hans Henrik Andersen
Statistics Denmark, Copenhagen, Denmark

Abstract. During the last 15 years labour market statistics produced by Statistics Denmark have increasingly become more integrated. For example, the Statistics on People Receiving Public Benefits have been joined into an integrated statistical system. In this way, the quality of the statistics has been enhanced and the published figures have become logically consistent. However, the statistical users request a more cohesive statistical system covering the entire population's attachment to the labour market. The system should include volume information and provide the possibility of analyzing longitudinal labour market data.

Against this background, in the beginning of 2012 Statistics Denmark initiated work on developing an integrated statistical system for analyzing the entire population's attachment to the labour market. The statistical system is called Labour Market Account (LMA). It is intended to publish statistics from the LMA in 2014. In addition to being an important source of the future labour market statistics, the LMA will also be a direct or indirect input source to a number of other statistics within social, business and economic statistics.

This presentation gives a description of the new statistical system and of the user requirements with regard to the system. Data from the various source registers frequently contain non-permissible overlaps or inconsistent start and end dates concerning the individual states. Subsequently, the presentation describes the core of the new statistical system which is the harmonization of data from a great variety of input sources and the longitudinal data processing conducted by the rule driven engine developed for this purpose.

Postscript. In April 2015 Statistics Denmark has published figures of the populations attachment to the labour market measured in full-time equivalents and as labour market status based on the new Labour Market Accounts (LMA). Longitudinal micro data for researcher will be available soon. The Working Time Account which provide data to the National Account will be revised in 2016 to use data from the Labour Market Account. The interest from users on the new possibilities is big.

Keywords: Integrated statistics, data harmonization, data linkage, micro data

Postscript. In April 2015 Statistics Denmark has published figures of the populations attachment to the labour market measured in full-time equivalents and as labour market status based on the new Labour Market Accounts (LMA). Longitudinal micro data for researcher will be available soon. The Working Time Account which provide data to the National Account will be revised in 2016 to use data from the Labour Market Account. The interest from users on the new possibilities is big.

1. Introduction

In 2012 Statistics Denmark began the work on developing a new integrated statistical system for labour market statistics. The integrated statistical system will make it possible to compile detailed structural statistics on the labour market in new ways.

In the new statistical system, micro data at individual level from various data sources are linked for the purpose of analyzing the population's labour market attachment (status and volume). In this context, a comprehensive data processing is performed, both of each data source separately and across the data sources. The new statistical system is called the LMA.

*Corresponding author: Pernille Stender, Statistics Denmark, Copenhagen, Denmark. E-mail: psd@dst.dk.

In the first part of the paper, a historic review is provided of the integrated register-based labour market statistics in Denmark. In the second part of the paper, it is explained how the LMA will be able to enrich the existing labour market statistics. The third part of the paper contains a description of the way in which the LMA is built-up and a description of the different types of processing to which the data are subjected. The focus will be on the establishment of input data sources, the rule engine and the major challenges in relation to the data processing.

2. The need for a LMA in Denmark – historical background

Statistics Denmark has a long tradition of developing integrated register-based statistical systems. The Register-based Labour Force Statistics was developed within the labour market statistics at the beginning of the 1980's. The statistics is compiled on the basis of an operationalization of the ILO's guidelines and show the Danish population's labour market status by the end of November each year. This implies that it is possible to conduct a socio-economic classification of the entire population at a detailed level. For persons in employment, the statistics is based on data linkages between each individual person and the establishment where the person works. Since 1981, the statistics has been published annually, and the statistics has been one of Statistics Denmark's most widely used register during the last 30 years.

As a result of the labour market policy measures, which were launched as a consequence of the economic crisis in the 1990's, there was a need for further development of the labour market statistics, enabling the statistics to provide a description of the new conditions in the labour market. Some of these needs were fulfilled by refining the Register-based Labour Force Statistics. Subsequently, new data sources concerning participants in the labour market policy measures were integrated in the statistics for the purpose of achieving a more detailed socio-economic classification.

As time passed, several separate statistics within the labour market statistics describing the development in employment and unemployment in full-time equivalents were developed. In some cases, these statistics contradicted each other with respect to levels and developments, which naturally gave rise to explanatory challenges for Statistics Denmark and interpretative problems for the users. Against this background, Statistics Denmark began by the end of the 1990's to prepare the theoretical frameworks for an integrated statistical system, known as the LMA, which could, among other things, analyze the population's labour market status in terms of full-time equivalents.

At the beginning, the LMA was intended as a statistical system based on micro data, but where the final product was aggregated data. The reason for this was that input data did not have the necessary content and adequacy to produce micro data statistics. It was especially a problem that the data quality concerning employee jobs was very low with respect to start and end dates for each job and with respect to volume information. This was also the reason why the Register-based Labour Force Statistics was only compiled once every year. It was only at end-November that the quality of the time reference of each employee job was acceptable. The plans for the LMA were never realized, but by the end of the 1990's Statistics Denmark published for the first time figures from the newly developed Working Time Account (WTA). The WTA is an integrated statistical system having common theoretical frameworks with the LMA, and producing statistics on employment, hours worked and total salary and wage costs intended for use in, e.g. the National Accounts.

In the 2000's two important data sources were developed, making it possible to redefine the LMA into a complex micro data based statistic. Also, the great need for longitudinal data, which was present among researchers and analysts, could now be accommodated.

The first important data source was the Statistics on People Receiving Public Benefits, which were integrated in a statistical system. In this statistical system, data are subject to processing of overlaps in the terms of time and volume, implying that a person can only be incorporated with 37 hours per week as maximum (corresponding to a full time standard). In addition to information on participation rates in each measure, the statistics also contains information on start and end dates. The Statistics on People Receiving Public Benefits contains information on persons, who are unemployed, in job activation, on early retirement pension, on early retirement pay, etc.

Secondly, Statistics Denmark gained access to a new administrative register (eIncome) containing monthly information on payments of wages and salaries and transfer incomes. Thereby, the problem of poor dates of employee jobs was to a large extent solved. In reporting data to the new register, employers were also, for the first time, to report data on the number of paid hours worked by each employee and far better volume information was thus achieved. A major development project was initiated by Statistics Denmark which purpose was to establish a register of employee jobs on the basis of the eIncome register. The register was to contain monthly information on job relations between persons and establishments. The data processing involved, e.g. selection of employee jobs, imputation of missing information on paid hours worked, enrichment of data with information on occupational status and placement of each individual employee at the correct establishment.

Statistics Denmark's new register containing employee jobs was fully developed by mid-2011. Thereby, Statistics Denmark now had high quality input data sources to the LMA. Consequently, the project of establishing a LMA was initiated at the beginning of 2012. The aim is to publish figures on the basis of the LMA at the end of 2014. The project is established as a Prince2 project, where the steering committee is composed of representatives from social statistics, business statistics, economic statistics and sales and marketing. The steering committee is chaired by the Director of social statistics.

The representatives from the three external departments in the steering committee have great interests in the LMA. With respect to the department for sales and marketing, the LMA will become a register that will open up unique possibilities for researchers and analysts of labour market statistics. In relation to the department for business statistics, the LMA will, e.g. become input source for the Business Register. With respect to the department for economic statistics, employment, jobs, hours worked and total wage and salary costs from the LMA via the WTA will be incorporated into the National Accounts. Also, the LMA can potentially become a valuable data input for the economic models.

3. New labour market statistics based on LMA

In the business case for the LMA the overall reasons for establishing the LMA are set out. The statistical system must, as something new for the labour market statistics, make it possible to.

3.1. Compile consistent statistics on the population's labour market status in terms of full-time equivalents

The purpose of the LMA is to give a full picture of the state in the labour market. A full picture of this type can only be given within the frameworks of an integrated statistical system, i.e. a statistical system with mutually consistent data. At present, statistics on the number of employees and persons receiving public benefits are compiled in terms of full-time persons. If these separate statistics are aggregated, the total will be too great, as there will be overlaps between the data sources. This is taken to mean that some persons may, e.g. have been registered as unemployed while at the same time they are in employment. Integrating the statistics in the LMA opens up the possibility of removing this overlap, and thereby achieving a more representative total for, e.g. the labour force.

3.2. Compile employment, jobs, number of establishments and total wage and salary costs

At the most detailed level the units in LMA are labour market spells. A labour market spell may be an employee job, a period of unemployment, a period of activation, etc. It is possible for the person to have several labour market spells at the same time. For all spells relating to jobs, there is a linkage between the person and the establishment and the person and the establishment both have unique identification codes. This implies that the number of jobs can be calculated. Furthermore the different spells which exit at the same time for a given person can be prioritized according to ILO-guidelines to calculated figures for employment.

It is also possible to sum up the states so the establishment becomes the statistical unit. LMA will also contain information on wages and salary costs for employees. This implies that personal as well as business statistics can be published on the basis of the LMA.

3.3. Develop more flexible classifications of the populations attachment to the labour market

There is a need for further knowledge concerning the different population groups which are not part of the core labour force. The ILO classification (ICSE) is unable to show whether a person has a duplicate state, i.e. whether the person appears in several different states at the same time. As the LMA contains all states for the person, it will be possible to analyze persons who have duplicate status (or are in intermediate forms in the labour market). The most common cases of duplicate states are as follows:

- Persons who are receiving education and at the same time have a job

- Persons receiving early retirement pay and at the same time have a job
- Persons who are pensioners and at the same time have a job
- Persons who are part-time unemployed and at the same time have a job
- Persons who are in subsidized employment

3.4. Describe longitudinal labour market statistics

The LMA will make it possible to publish numerous types of dynamic statistics. Two possible ways of showing the dynamics in the labour market is to compile gross flows between two status points-in-time or by preparing dynamic classifications.

Gross flows describe the movements between reference points-in-time. Gross flows will typically be compiled in relation to the primary labour market status of the population. Hitherto, it has only been possible to compile gross flows between points-in-time in November, because the Register-based Labour Force Statistics were related to end-November. When LMA is developed, the period of time in which the gross flow takes place can be chosen optionally. One type of gross flow may be gross flows among the various socio-economic groups. Another type may be gross flows for persons in labour market policy schemes at one point in time and their socio-economic status at another point in time.

Another possibility will be to take specific populations as starting point and subsequently estimate the average time until these populations switch to another state. This would typically be relevant in connection with the transition from the education system to the labour market.

Dynamic classifications are classifications where the population's socio-economic status has been determined on the basis of information covering a lengthy period of time. There are no existing international standards for dynamic classifications, but the ILO has some years ago put forward a proposal for a classification. With the establishment of the LMA it is natural to prepare dynamic classifications of the population in relation to the labour market status of the population.

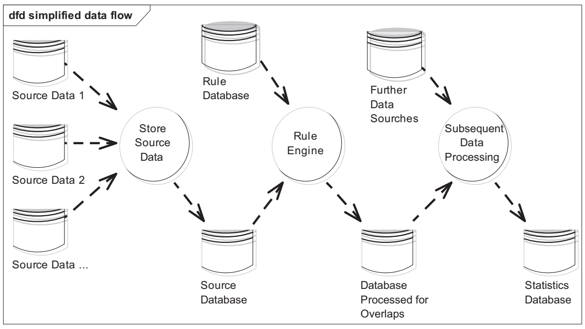
3.5. Conduct status observations of the population's labour market status at arbitrary points-in-time during the year, and enable the calculation of average figures for a given period

Hitherto it has only been possible to calculate the attachment of the population to the labour market at the end of November each year. When LMA is developed it will be possible to calculate the attachment to the labour market at arbitrary points in time during the year. It will also be possible to calculate e.g. the average employment during a period e.g. a year. This implies that employment is calculated on a daily basis and thereafter calculated as an average over the period in time.

4. Set-up of the LMA

The data processing in the LMA will be conducted by the steps below:

- Data processing of the individual data sources and storing of data in a source database
- Processing of data between data sources subject to overlaps. This work is carried out by the rule engine. The dynamic rules are stored in a rule database. Data processed for overlaps are stored in the database processed for overlaps
- Compressing, summation and classification. Linkages to other registers, final data editing, if required, and storing of data in a statistics database



4.1. Source database for the LMA

The LMA is flexible with respect to the selection of data sources. This implies, in practice, that the system is established so data sources can be replaced, when new or better data possibilities are available.

During the period January 2012 to June 2013 a major part of the work was devoted to analyzing which data sources are to be used by the LMA. This includes specifying which kind of checks are to be carried out in order to reveal errors and problems of consistency and how the data are to be edited before data can be stored in the source database. The data editing is implemented using a source-specific program that will perform the necessary data editing and harmonization of that particular source. In addition to this, there is a common program, ensuring that the source data are subsequently stored in a standardized form in the source database. This program defines and handles, e.g. how rows of data are entered or deleted in the source database. The overall premise is that only new, deleted or altered rows are recorded in the database.

At the end of June 2013 the LMA source database contained information from all source registers relevant at this stage of the LMA. The database includes information on:

- Unemployment, labour market policy measures, early retirement and certain other social benefits. These data are based on the Statistics for People Receiving Public Benefits, containing data that have been processed for overlaps. An analysis has revealed that the processing of time and volume overlaps and the prioritizing of data sources are not problematic in respect to the LMA[1] specifications with one exception only: People temporarily absent from the labour market because of maternity or sickness leave. Such states are sometimes overruled in the processing of overlaps carried out. In the LMA, temporary absence must be identified. As a consequence, source data regarding maternity and sickness benefits are based on data from before processing of overlaps in the Statistics for People Receiving Public Benefits.
- Employee jobs. These data are based on the Quarterly Employee Statistics. In this register, eIncome (micro level) data are further edited regarding, e.g. the linkage of jobs to establishments. A separate version of the register has been created for the LMA to include late reporting not included in the Quarterly Employee Statistics and to improve quality even further. The tasks conducted to improve the data quality even further – both with regard to the Quarterly Employee Statistics and the separate version for LMA – have been relative comprehensive. These improvements often deal with data problems relating to specific groups of employees at a given time. Such data problems are typically not solved when the eIncome register is formed. An example is a situation where a unique identifier for an enterprise is missing at a given point in time. This implies that information about skill level from earnings statistics is not linked to employees at the enterprise.
- Jobs by self-employed and assisting spouses. These data are based on a variety of sources. There are four main challenges relating to data on self-employed and assisting spouses: Firstly, the information on start dates and end dates of the jobs is relatively uncertain. Secondly, it is uncertain how the activities are distributed across a given period of time. Thirdly, the activities are in many cases very small and it is therefore doubtful whether it really is a job. However it typically cannot be ruled out in advance that it may not reflect a job. Fourthly, there is no administrative information on hours worked. In the formation of the data sources for this group, data are compared at the individual level with information on persons included in the Labour Force Survey (LFS). On the basis of these analyses, persons with various characteristics in the LMA source data[2] and LFS information are joined and a likelihood is established for the jobs to reveal an actual activity as self-employed in the reference period.

  In the formation of source data an imputation of hours worked by self-employed and assisting spouses is also conducted. The imputation is conducted on basis of:

  ∗ Information on paid hours of work for employees in their primary job in the sector group 'corporations and organizations' (from the Quarterly Employee Statistics)[3]

  ∗ The ratio between actual and paid hours of work for employees in the Earnings Statistics (to convert from paid hours of work to actual hours worked)[4]

  ∗ The ratio between actual hours of work by self-employed and assisting spouses in the LFS compared to actual hours of work by employees in the LFS (to correct for the fact that self-employed typically work more hours than employees).

[1] In LMA the guidelines for the International Classification of Status in Employment are to be followed.

[2] The characteristics could be, e.g. purchases and sales by firms ("VAT statistics"), surpluses for self-employed in the year, whether the self-employed is employing any employees, or whether the person is insured in an unemployment fund for self-employed.

[3] The background variables used are detailed line of industry, age and gender (the choice of line of industry being the most detailed level ensuring at least 200 jobs in the cell as basis for the calculation).

[4] These ratios are conducted annually and broken down by aggregate line of industry (8 groups) and gender.

  - ∗ The same information is used when hours worked in sideline jobs for self-employed are estimated, although this calculation include information from the Quarterly Employee Statistics on the ratio between paid hours of work in sideline jobs compared to paid hours of work in primary jobs for employees.
  - ∗ When hours worked are distributed over the course of the year, information on the distribution of monthly hours of work per job over the year (from employees in the LFS) relative to the distribution of monthly paid hours of work per job over the year (from the Quarterly Employee Statistics) are used along with information on Easter Holiday effects and leap year effects in the year.
- Persons receiving maternity/paternity or sickness benefits, i.e. persons temporarily absent from the labour market. These data are based on data from the Statistical Register for Persons Receiving Maternity/Sickness benefits. In this register, an algorithm has been developed which – with a varying degree of certainty – determines if the person in question is absent from employment or from unemployment. This is necessary because there are situations where there are no payments of wages or salaries when a person is absent due to maternity/paternity leave. Nevertheless the person should be considered as employed due to international guidelines.

  When establishing source data for persons receiving maternity/paternity or sickness benefits it was decided to introduce a new division of work between the donor register (the Statistical Register for Persons Receiving Maternity/Paternity or Sickness benefits) and LMA. This implies that a preliminary estimation concerning whether the person is absent from employment or unemployment is conducted when forming the donor register. The estimation is solely based on information which is reported to the donor register. Results from the estimation are carried forward as a data input to the LMA. Thereafter the final estimation is made by LMA. In the final estimation information from other data sources available in the source database for LMA are also used, e.g. information concerning jobs at other points in time for the particular person. When it is estimated that the person is most probably absent from employment a job is imputed in situations where there are no payments of wages or salaries, but the persons are still employed. Finally, information about whether the person is absent from employment or not is delivered from LMA to the Statistical Register for Persons Receiving Maternity/Paternity or Sickness benefits. This information is used when publishing statistics based on the donor register.
- Participants in education. These data are based on the Register of Educational Statistics which is a longitudinal register.

In the Register-based Labour Force Statistics there is a major group of the population which cannot be assigned any socio-economic status. In connection to the formation of the LMA source data, analysis has been carried out with the aim of reducing the size of this group. The analysis has gained only limited results, but nevertheless it has been revealed that data on participants in educational courses and students receiving public grants can reduce the size of the group. Therefore, these data and also data on retirement benefits will be used in the LMA. As these data are not to be used in the processing of overlaps, they will be added at a later stage in the data processing. More work has to be carried out with the aim of getting more knowledge about the group which cannot be assigned any socio-economic status.

4.2. The rule engine and entry of rules

The "rules engine" is the popularized name for the framework-system in the processing of overlaps. The rule engine ensures that overlapping source data are processed according to a set of specified rules. The rules serve the purpose of:

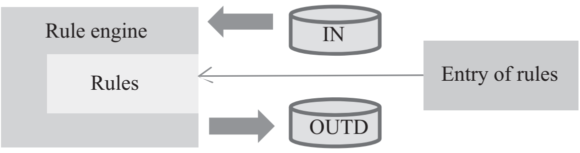
- Deleting erroneous states
- Reducing the number of hours in one or more states when the states of each person is summing to a total of more than the allowed number of hours (the full-time standard)
- Editing information on start- and end-dates for better longitudinal coherence

The rule engine has been constructed according to the same principles as the rule engine used in the Statistics for People Receiving Public Benefits. Nevertheless it has been necessary to modify the rule engine because LMA has a different data model and also LMA has other requests in relation to functionality.

The framework-system (rules engine) contains:

- Parameterization of which processing of overlaps is implemented.

- Reading of input data from the source database that the processing of overlaps indicates.
- Writing of output data to the databases the processing of overlaps.
- Split-up of time references of all data, implying that data can be evaluated for periods of time instead of individual single dates.
- Permutation of all data, resulting in a simplification of laying down rules.
- Possibility of both parallel and sequential evaluation of the rules.
- Parallel implementation of data subject to processing of overlaps
- Batch implementation of data subject to processing of overlaps.
- Counting included in the log and sent by e-mail to the operator.



Furthermore, the processing of overlaps comprises data and code components, describing the rules entered by the developers of rules. The entry of the rules is performed via a system that validates and prepares the rules as part of the optimization of the data subject to processing of overlaps.

The entry of rules comprises:

- Grouping of rules into a compendium of rules and linkage of the compendium of rules for processing of overlaps.
- Establishment, correction and deletion of rules
- Automatized programming generating of the part implementing the rules during the processing of overlaps.

The rules for subjecting data to processing of overlaps in the LMA can be divided into two types of rules called, respectively, static and dynamic rules. Static rules can be more complex than the dynamic rules, e.g. they can change starting and end dates of the spells. Static rules must always be implemented by a programmer. The dynamic rules are based on stable periods of time where all states for a person are the same during the entire period. The components implementing dynamic rules are generated on the basis of the rules database and consequently dynamic rules can be specified solely by the user.

4.3. Processing of overlaps across sources – specification of rules

When the complex rules are set up handling overlapping spells from various sources, the quality of the dates and hourly information must be taken into account. Another criterion to take into account is whether the different spells are expected to, or expected not to, exist simultaneously. An example of spells that are expected to exist simultaneously is a spell concerning subsidized employment in the Statistics for People Receiving Public Benefits and a spell concerning a job in Quarterly Employee Statistics. An example of spells which are not expected to exist simultaneously is a job spell where the person is working full-time and an unemployment spell. For the specification of rules, a number of analysis programs have been developed, mapping overlaps between different states and calculating consequences of implementing different rules for the processing of overlaps.

The processing of overlaps is the core of the LMA. Analyzing overlaps and setting up rules for the data processing have been initiated by the end of June 2013 and are expected to last until the end of 2014. The processing will be based on a weekly standard of 37 hours, so that the sum of hours for one person will not exceed this norm at any time. In addition, an alternative set of rules not limited to the 37-hour norm will be defined in order to take into account the needs of the WTA, being a future user of data from the LMA.[5]

At present, work has been conducted with the aim of analyzing and specifying rules for the following groups:

- Self-employed and assisting spouses
- Persons in subsidized employment
- Persons receiving maternity and sickness benefits

Self-employed and assisting spouses are by far the group for whom the smallest item of information is available and information is typically highly inadequate. Firstly, there is no register-based information as to the number of hours worked by these groups and secondly, the information as to when the jobs are active is subject to great uncertainty. That implies that the current figures for hours work for self-employed and assisting spouses in in the Working Time Account and

[5] The WTA produce information on, e.g. hours actually worked for the National Accounts as a measure of labour input into the production process. All hours worked, including those beyond 37 hours a week, have contributed to the production and must therefore be included in the WTA.

the National Account is subject to a higher degree of uncertainty compared to hours worked for employees.

Integrating information about other activities than employment for self-employed and assisting spouses in the LMA will undoubtedly improve this to a great extent. At the moment, different methods for data processing of overlaps of self-employed and assisting spouses are been tested. Analytical work is still being carried out, and in this connection the Labour Force Survey is used.

Information on labour market spells for persons in subsidized employment is presently available in three of the data sources for the LMA. These sources comprise the Statistics for People Receiving Public Benefits, the Quarterly Employee Statistics and data for self-employed. During the data processing these spells of events must be linked, implying that the sequences of events in the Statistics for People Receiving Public Benefits are linked to the relevant job in the other two data sources. In many cases, the dates in the input data will not be exactly the same. The number of hours may also differ. Consequently, analyses have been conducted for the purpose of determining which of the data sources are best suited as basis for information on the number of hours worked, depending on the employment measure. Furthermore, rules with regard to harmonizing to the start- and end-dates across the input data sources have been worked out

For persons receiving maternity and sickness benefits the work has primarily been concentrated on implementing the rules adopted from the donor register (cf. p. 7) and secondly, analyzing and improving the rules with regard to the possibilities which are now available using information from the harmonized source database for LMA. Another focus point has been how the division of labour between the donor registers should be carried out in practice. Here one of the challenges has been that the two statistical fields have different demands with regard to how compressed data should be.

4.4. *Subsequent data processing*

After the data have been processed for overlaps, further data sources will be added with the aim of working out a detailed socio-economic classification of persons outside the labour force or in order to define the population.

On the basis of the data processed, one or more statistics registers must be defined. These registers have not yet been specified. The objective is that the registers are flexible, in order to allow for, e.g. different socio-economic classifications of the population. These classifications may be either static or dynamic.

Of course, the LMA contains unique identifications of the individual person and establishment. By the use of these identifications, background information on persons and establishments can be found in other registers. Information on persons, such as residential address, education and ancestry, is gathered from a database in the social statistics, where standardized versions of social statistics are stored. Information on establishments, e.g. address, line of industry and sector information, is gathered from frozen versions of the Business Register ensuring that the various statistics are as comparable as possible.

5. Conclusion

With the development of the LMA, Statistics Denmark will have a micro data register making it possible to describe the population's labour market attachment defined in various ways with detailed background information. In addition to measuring status and volume, it will also be possible to analyze flows between different groupings, i.e. longitudinal analyses. It can be used as a direct source for publishing statistics as well as a tool for improving the quality of other statistics and registers. Furthermore, the LMA will enable Statistics Denmark to produce tailor-made solutions for its customers according to their specific needs. Finally, micro data can be offered to the research community. Consequently, the LMA will make important contributions to the description of the population, the labour market and the economy.

At the present state the developing work is in progress according to the plan and we are looking forward to publish the first results based on LMA at the end of 2014.