7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

Opening session

Introducing modularity in order to improve quality and efficiency

Johan Van der Valk – Eurostat

COMMISSION OF THE EUROPEAN COMMUNITIES
EUROSTAT

Directorate F: Social statistics
Unit F-3 :Labour market

7th Workshop on LFS methodology, Madrid, 10-11 May 2012

Introducing modularity in order to improve quality and efficiency[1]

Johan van der Valk (Eurostat)

Summary

In this document it is argued that modularising complex surveys like the Labour Force Survey contributes to improve the quality of the data and increase the efficiency of the production of data. A modular architecture structures and simplifies the data collection system. Moreover, it enables to link the required output on one hand and the collection and processing of the data on the other hand. The adaptations required for labour market variables to introduce modularity are limited. Once this architecture is established it paves the way to incrementally improve the production of the statistics based on surveys.

1. Introduction

This document presents ideas about the role that modularity could play to improve surveys. The Labour Force Survey (LFS) is used to see the practical implications of introducing modularity for a complex survey. Modularity plays a crucial role in the project of streamlining social surveys to modernise EU social statistics. In this context an architecture for a new system of social statistics is currently being developed. The aim of this document is to obtain feedback to be used as input for the process of reviewing the LFS and the modernisation of social statistics.

2. What could modularity be in case of social surveys?

Searching the Internet one gets a good idea what modularity means. Modularity[2] is a general systems concept, defined as a continuum describing the degree to which a system's components may be separated and recombined. It refers to both the tightness of coupling between components, and the degree to which the rules of the system architecture enable or prohibit the mixing and matching of components. In software designing, modularity is a logical partitioning of the software design that allows complex software to be manageable for the purpose of implementation and maintenance.

Modules have the following essential characteristics. Firstly, they are domain specific. They are specialised responding to input or perform functions to a certain group of objects. Secondly, they are to a high extent autonomous. This means that there is almost no interdependence between modules. There may be some interaction between modules, but the

[1] The views expressed in this paper are those of the author and do not necessarily reflect the opinions or policies of Eurostat.

[2] http://en.wikipedia.org/wiki/Modularity

greater interaction and integration occurs within the module. Thirdly, they can be hierarchically nested. This means that each module can be decomposed into finer modules.

In social statistics a module could be defined as a set of variables homogeneous in terms of content measuring a topic comprehensively. The set should be such that there is a strong association of the variables within the module and very weak, preferably non-existing, relations between modules. Filters for the modules defining the target group should therefore consist of a minimal number of elementary variables only. On the other hand, filters for variables are determined by other variables in the module. Hierarchically nesting implies that modules could be further decomposed into sub-modules using the same reasoning on a lower level.

A module is primarily defined by the required output that it should provide. Per module needs to be specified what should be measured, why, when, and with which quality. More concretely has to be specified: the target group, the set of variables, the objectives, the reference period of the data collection or frequency and the required quality. The latter element can involve several aspects but the main aspect will be the accuracy that could be translated into a sample size. The set of variables is a crucial element of the definition. It should be a homogenous of contents and not strongly related to other modules.[3]

In order to have the full advantage of modularity, the specified output of a module should be linked to the input of the data collection. This means that the modules should be translated into questionnaire modules. Technically this is no problem in case of computer assisted data collection modes. In these modes the questionnaire is in fact a software program. Modularity in software design is commonly used. Therefore one-to-one correspondence is quite straightforward. The characteristics of the modules to be independent of other modules and that the contents is homogenous facilitates the translation into a questionnaire module.

Once modules are developed they can be used to conduct surveys. A survey is generally defined as a method used to collect in a systematic way information from a sample of individuals. Such a survey can be designed as a set of modules, measured in a certain period for a certain population. Every survey requires including modules on basic background information and a module with technical information. The actual survey then is defined by the set of contents matter modules.

3. What are the advantages in applying modularity?

A main advantage of modularity is to give structure to a complicated system. Currently, surveys are defined by a long list of variables with filters. It is difficult to have good overview of the whole list and the relationships between variables. Modules consist of a set of variables on one topic for a clear target population. This makes it easy to have an overview. Moreover, a set of more or less independent modules are defined with limited in-between relations simplifies the system considerably. Currently, filters are defined per variable. They can be complex depending on a number of other variables. Defining filters per variable makes the relations between variables numerous. Limited and straightforward relations between modules simplify the allowable interrelationships between variables. It reduces the degree of freedom somewhat but it increases the logical structure.

Similarly to the use in software design modularity facilitates complex data collection systems to be manageable for the purpose of development, implementation and maintenance. These

[3] Several LFS ad hoc modules do not meet these requirements and can therefore not be considered modules in sense of what is proposed here.

activities are carried out on the level of the module instead of the current traditional surveys. Instead of reviewing and improving whole surveys this can be done per module.

Modular design of surveys allows a more efficient production of social statistics. Modules are developed and maintained by content matter specialists. It can be used for all social statistical data sets for which a particular content is needed. This would avoid double work because the module is developed once and not for each data collection round. In addition, harmonisation of the output for data sets is more or less guaranteed. Furthermore, modules can be specified and directly linked to meta-databases independently of the data collection rounds.

Modularity makes it possible to better fit the data collection to output needs. Currently, frequencies and timings are determined by the design and data collection period of a whole survey. As a consequence, all variables are measured for the same sample, with the same frequency at the same moment. However, the output requirements differ per set of variables. Modularity makes it possible to include or exclude a module easily in a data collection period. Modules can have different timings. Data can be collected multi-annually instead of annually or only in one quarter instead of quarterly for a specific module. Modules can be applied to subsamples or sub-populations. Modules can even be implemented as follow-up surveys with a specific data collection mode. This can all be tailored to the output and data requirements.

Modular design of questionnaires facilitates maintenance of the questionnaires and the data processing activities. Theoretically, modules can be processed independently. In that case a change to one module will not affect the others. This is effective and reduces risks of errors.

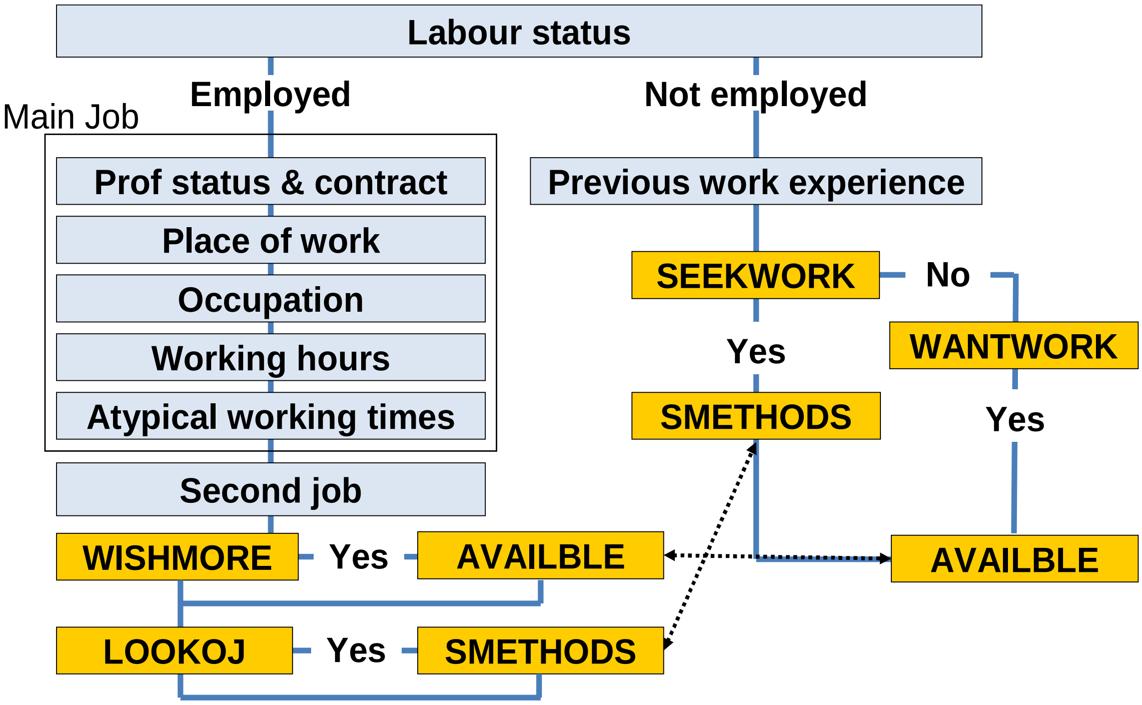
4. Putting modularity in practice: the case of the EU LFS

To see what impact it would have to introduce modularity for complex surveys the EU LFS is analysed. The current LFS codification list is taken as a basis. Subsequently problematic elements are identified to see which changes are necessary and what consequences this would have. To simplify the analysis the exercise was limited to the labour market variables only.

The headings in the current codification list of the LFS cannot be considered modules for several reasons. The labour market variables are listed under the following headings: 'Labour status', 'Employment characteristics of the main job', 'Atypical work', 'Hours worked', 'Second job', 'Previous work experience of person not in employment', 'Search for employment', 'Methods used during previous four weeks to find work'. First of all, the current groupings are quite arbitrary. The heading of 'Employment characteristics of the main job' can contain almost anything while there are separate headings for 'Search for employment' and 'Search methods'. Secondly, the labels do not reflect the contents very well. For example, the label 'Employment characteristics of the main job' suggests that the headings atypical work and hours worked do not refer to the main job. The heading 'Hours worked' contains variables on looking for another job and how the job was found which would be expected by a user. On the other hand part-time work is not part of this grouping but is included under the heading 'Employment characteristics of the main job'. Finally, the groupings are not independent. There are too many variables used as filters for variables under other headings. In order to group variables in modules some changes are necessary. Labels of the modules must be self-explanatory, contents homogeneous, size of groups balanced and interdependencies removed. This can be realised by moving variables, changing filters and splitting variables.

A possible approach with suggestions how to adapt the list is worked out in detail. A set of labour market modules always starts with a basic module on labour status to distinguish two target populations: employed and not employed. Subsequently, for employed characteristics of their jobs and their wishes and activities to change jobs are measured. Several modules are required to collect information on the main job. A possible set of modules is: professional status, characteristics workplace, occupation, working hours, start work and atypical work. This part of the decomposition is relatively straightforward. Order, contents and size of modules can be adapted but this is not essential. The only issue here is to what extent one would like to allow that the distinction employee/self-employed should play a role in the other modules. For the not-employed persons having a module on previous work experience is straightforward. This is not the case for the variables on looking for work and availability. To put them into modules is problematic. This is caused by the fact that these variables are applicable to both employed and not-employed and secondly rely on both looking for work and on hours worked variables. This entangles the variables in such a way that decomposition in modules is not really possible. For employed is measured if they want to work more hours and if so if they are available. This is required to identify underemployed persons. In addition, is measured if they are looking for work and which methods they use. For the not employed these search methods are measured if they are looking for work and availability is measured if they seek for work or would like to work. A schematic representation in scheme 1 with the main variables shows the complicated structure. As a result countries have to design complicated questionnaires with strange and difficult routings.

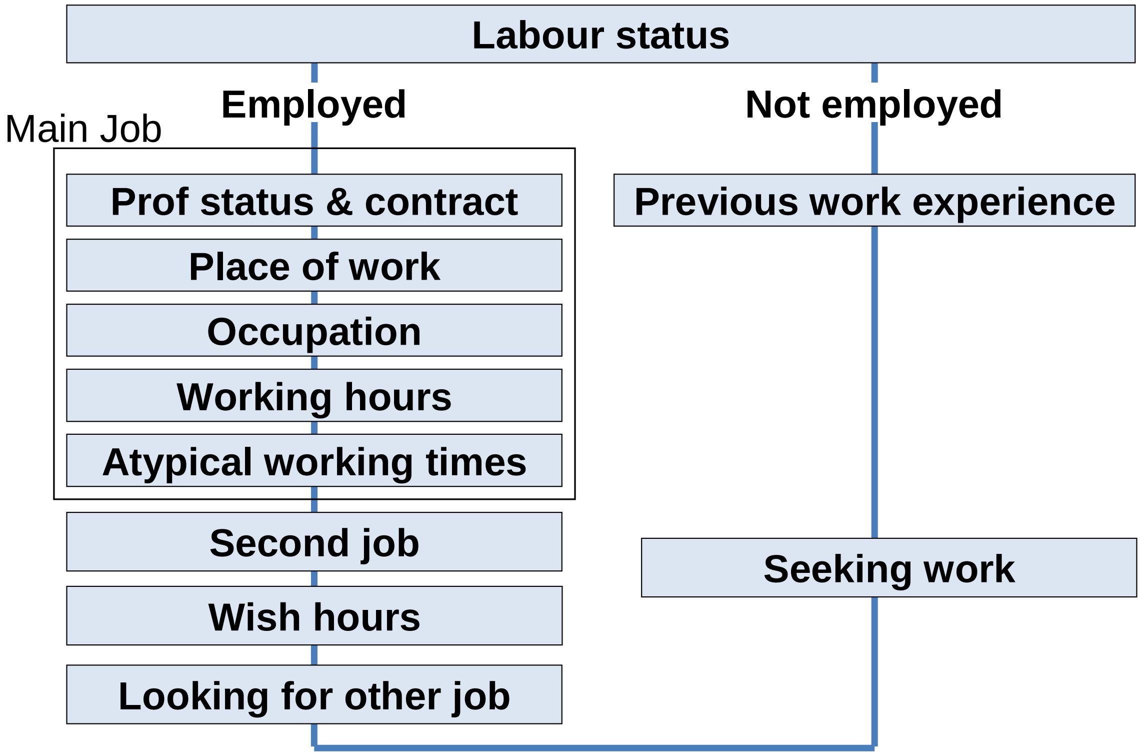
Scheme1. Flowchart of possible labour market modules and variables in the current situation



To be able to make modules the variables currently combined for both employed and not-employed should be split. For both search activities and availability two separate variables with simple filters can be defined. This allows a logical decomposition into modules. To simplify the matter the variables on job search methods for employed one could consider deleting these variables. They seem not essential and are not required to measure underemployment. In that case only availability in scheme 1 should be split into two variables. From scheme 1 is clear that it will have to be asked in different positions in the questionnaire anyway. Moreover, splitting the availability variable in two separate variables also better reflects the fact that the variables are measuring two different concepts. Being available to start working if a person has no work is different from being available to work more hours. This can be illustrated with the UK questionnaire 2010. It contains different questions to these populations: UNDST for employed and START for not employed looking for work. Employed that would like to work more hours in the current job are asked: '*If you were offered longer hours (by your current employer) could you start working them within two weeks?*'. Other employed that would like to work more hours in the current job are asked: '*If you found a job or work to provide these extra hours could you start within two weeks?*'. Persons not at work looking for work are asked: '*If a job or a place on a government scheme had been available in the week ending Sunday the [date], would you have been able to start with two weeks?*'. This split into two questions is also applied in more countries than the UK. Thus splitting the variable is more consistent the actual measurement.

Once the variable is split separate modules for employed and not employed can be defined making the structure much more logical. For the employed two modules can be defined: Wish hours and Looking for other job including the variables AVAILHRS and LOOKOJ. For the not-employed one module on Search for employment including search methods and availability is sufficient. To make this possible it is also required to split the variables NEEDCARE and REGISTER. In scheme 2 the resulting simple decomposition is presented.

Scheme 2. Flowchart of possible labour market modules and variables in the new situation



Annex 1 presents a preliminary possible way of defining LFS labour market modules. Some variables were moved, split and filters adapted. Filters of modules define the target population. Since this is valid for all variables in the modules it does not have to be repeated. To signify this simplification the filters relating only to the target population of the module are removed. In most cases the remaining filters rely on variables in the module only. This implies that the modules are quite independent. The remaining relations between modules are those that cannot be avoided. Labour status has to precede all other labour market modules. Professional status precedes all other main job modules. Working hours and (the hours variable of the) second job needs to precede the module Wish hours.

The suggested way of modularisation does not affect the output. In the proposal only some variables on job search were removed because they do not seem essential. However, if they are deemed to be too important to drop they could be included without changing the architecture.

5. Conclusions and final remarks

The specifications of labour market variables as currently is laid down in the LFS legal basis needs to be adapted in order to define labour market modules. By moving, splitting and widening the filters modules and sub-modules can be designed. All in all, the required adaptations of the labour market variables are limited while the output stays virtually the same. This means that the costs of introducing such architecture are therefore low.

The potential gains are large. The modular architecture can be used to improve the quality and increase the efficiency of data collection. For instance, per module can be looked at the contents to see if it requires improvement. For modules with high quality requirements could be considered to define a model questionnaire. This is currently done for the labour status module and the seeking work module in the context of the task force on the measurement of labour status. The module on working hours seems also a good candidate for such an exercise.

Furthermore, if full consistency between modules and data collection units or questionnaire is ensured data processing can be organised accordingly. Modules can be managed as separate units. This can applied for development and maintenance in all phases of the measurement process until the dissemination phase.

The modular architecture can be also used to substantially increase the efficiency of the LFS data collection. However, severe measures are required to realise this. Per module, the required sample sizes and the frequency should be critically assessed. Currently, only ad hoc modules and atypical working times are annual modules which allow applying sub-sampling. One can question if for all other modules a quarterly frequency is required. For modules like occupation, characteristics of the workplace, start work, second job, wish hours and looking for another job an annual frequencies can also be considered. This issue will have to be addressed applying the appropriate procedures and a long timeline. The nice feature of modularity is that it is possible to take one step at the time handling module per module.

Once modules are defined with different frequencies they can be included in a LFS system in a natural way. Modules can be assigned to a specific wave in order really introduce the wave approach. The first or the last wave are logical candidates but other waves are also possible.

In addition, when labour market modules are defined they can be used for all national or international data collections where such variables are required for target or background variables. It can be seen as a sort of menu where users can select from. Including a module in a computer assisted environment is very easy if all elements of the process are available. This would consist of a questionnaire in the form of a CAWI or CATI programme, data processing programmes related to this module and the meta-data. If the whole package is available per module they can be included effortlessly.

Annex 1. Possible assignment of variables to labour market modules

LABOUR STATUS			**Age>14**
	WSTATOR	Labour status during the reference week	
	NOWKREAS	Reason for not having worked at all though having a job	WSTATOR=2
M	SIGNISAL?	Continuing receipt of the wage or salary	(WSTATOR = 2 and NOWKREAS ≠04 and NOWKREAS≠ 05) or WSTATOR= 3

If working:

MAINJOB			**WSTATOR=1,2**
PROFESSIONAL STATUS AND CONTRACT			
	STAPRO	Professional status	
	TEMP	Permanency of the job	STAPRO=3
	TEMPREAS	Reasons for having a temporary job/work contract of limited duration	TEMP =2
	TEMPDUR	Total duration of temporary job or work contract of limited duration	TEMP=2
	TEMPAGCY	Contract with a temporary employment agency	STAPRO=3
CHARACTERISTICS WORKPLACE			
	NACE3D	Economic activity of the local unit	
	SIZEFIRM	Number of persons working at the local unit	STAPRO=1, 3, 4, Blank?
	COUNTRYW	Country of place of work	
	REGIONW	Region of place of work	
M	HOMEWK	Working at home	
OCCUPATION			
	ISCO4D	Occupation	
	SUPVISOR	Supervisory responsibilities	STAPRO = 3?
WORKING HOURS			
	HWUSUAL	Number of hours per week usually worked in the main job	
	HWACTUAL	Number of hours actually worked in reference week in the main job	
M	FTPT	Full-time /Part-time distinction	
M	FTPTREAS	Reasons for the part-time work	FTPT=2
N	NEEDCARW	Need for care facilities for working PT	FTPTREAS=3
	HWOVERP	Paid overtime in the reference week in the main job	STAPRO=3?
	HWOVERPU	Unpaid overtime in the reference week in the main job	STAPRO=3?
	HOURREAS	Main reason for hours actually worked during the reference week being different from the person's usual hours	HWUSUAL=00-98 & HWACTUAL=00-98 & WSTATOR=1
STARTWORK			
	YSTARTWK	Year in which person started working	
	MSTARTWK	Month in which person started working	YSTARTWK≠9999, blank & REFYEAR-YSTARTWK<=2
C	WAYJFOUN	Involvement of the PEO in finding the present job	has started this job in the last 12 months
ATYPICAL WORKING TIMES			
C	SHIFTWK	Shift work	STAPRO=3?
	EVENWK	Evening work	
	NIGHTWK	Night work	
	SATWK	Saturday work	
	SUNWK	Sunday work	
SECOND JOB			**WSTATOR=1,2**
	EXIST2J	Existence of more than one job or business	
	STAPRO2J	Professional status (in the second job)	EXIST2J=2
	NACE2J2D	Economic activity of the local unit (in the second job)	EXIST2J=2
	HWACTUA2	Number of hours actually worked during in the second job	EXIST2J=2

N=New, M=moved, C=changed

WISH HOURS			**WSTATOR=1,2**
	WISHMORE	Wish to work usually more than the current number of hours	
	WAYMORE	Way how person wants to work more hours	WISHMORE=1
	HWWISH	Number of hours that the person would like to work in total	
N	AVAILHRS	Available to work more hours	WISHMORE=1
N	AVHRREAS?	Reasons for not being available to work more hours in 2 wks	

LOOKING FOR OTHER JOB			**WSTATOR=1,2**
	LOOKOJ	Looking for another job	
	LOOKREAS	Reasons for looking for another job	LOOKOJ = 1
N	REGISTERW	Registration at a public employment office to find other job	

If not working:

PREVIOUS WORK EXPERIENCE OF PERSON NOT IN EMPLOYMENT			**WSTATOR=3-5**
	EXISTPR	Existence of previous employment experience	
	YEARPR	Year in which person last worked	EXISTPR=1
	MONTHPR	Month in which person last worked	YEARPR≠9999, blank & REFYEAR-YEARPR <= 2
	LEAVREAS	Main reason for leaving last job or business	EXISTPR=1 and REFYEAR -YEARPR<8
	STAPROPR	Professional status in last job	EXISTPR=1 and REFYEAR -YEARPR<8
	NACEPR2D	Economic activity of the local unit in which person last worked	EXISTPR=1 and REFYEAR -YEARPR<8
	ISCOPR3D	Occupation of last job	EXISTPR=1 and REFYEAR -YEARPR<8

SEEKING WORK			**(WSTATOR=3-5 or SIGNISAL=3?) and Age<75**
	SEEKWORK	Seeking employment during previous four weeks	
	PRESEEK	Situation immediately before person started to seek employment	SEEKWORK=1, 2, 4
	SEEKREAS	Reasons for not seeking employment	SEEKWORK=3
M	NEEDCARE	Need for care facilities for not working	SEEKREAS =3
	WANTWORK	Willingness to work for person not seeking employment	SEEKWORK=3
C	SEEKTYPE	Type of employment sought (or found)	SEEKWORK=1, 2, 4
C	SEEKDUR	Duration of search for employment	SEEKWORK=1, 4
C	METHODA	Contacted public employment office to find work	SEEKWORK=4
C	METHODB	Contacted private employment agency to find work	SEEKWORK=4
C	METHODC	Applied to employers directly	SEEKWORK=4
C	METHODD	Asked friends, relatives, trade unions, etc.	SEEKWORK=4
C	METHODE	Inserted or answered advertisements in newspapers or journals	SEEKWORK=4
C	METHODF	Studied advertisements in newspapers or journals	SEEKWORK=4
C	METHODG	Took a test, interview or examination	SEEKWORK=4
C	METHODH	Looked for land, premises or equipment	SEEKWORK=4
C	METHODI	Looked for permits, licences, financial resources	SEEKWORK=4
C	METHODJ	Awaiting the results of an application for a job	SEEKWORK=4
C	METHODK	Waiting for a call from a public employment office	SEEKWORK=4
C	METHODL	Awaiting the results of competition for recruitment to public sector	SEEKWORK=4
C	METHODM	Other method used	SEEKWORK=4
C	AVAILBLE	Availability to start working within two weeks	SEEKWORK=1, 4 or WANTWORK=1,blank
	AVAIREAS	Reasons for not being available to start working in 2 wks	AVAILBLE=2
M	REGISTER	Registration at a public employment office	