Distr. GENERAL
19 September 2013
WP 2
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on Statistical Data Collection
(Geneva, Switzerland, 25-27 September 2013)

Topic (i): Centralising data collection

LINKS BETWEEN CENTRALISATION OF DATA COLLECTION AND SURVEY INTEGRATION IN THE CONTEXT OF THE INDUSTRIALISATION OF STATISTICAL PRODUCTION

Working Paper

Prepared by Fernando Reis, European Commission (Eurostat)[1]

I. Introduction

1. An important part of the efficiency gains we can get in the data collection process comes from the way we design our surveys. Even with the centralisation of the data collection, continuing to design the surveys as stand-alone entities ignoring that, in the context of official statistics, they are run together with many other surveys, limits the potential efficiency gains. The centralisation of the data collection depends on, and supports, the integration of the surveys.

2. In the context of the modernisation of social statistics, Eurostat has been working on the streamlining and integration of the European social surveys. The approach adopted was to move towards the adoption of a modular architecture of social surveys. The organisation of the content of the surveys in modules is not new in the social surveys. However, the adoption of a modular architecture requires the development of modules with particular characteristics. This type of architecture is based on the re-usability of the modules in the development of different surveys. It impacts all the steps of statistical production. However, its main potential efficiency gain comes from the impact it is expected to have on the data collection process. The centralisation of the data collection is a requirement for the adoption of the modular survey architecture. But this architecture also requires the data collection process to be much more flexible and ready to deal with higher complexity in terms of which data to collect from which respondent at which time.

3. This paper presents the experience so far in Eurostat on the modularisation of the European social surveys and in investigating how an integrated system of surveys based on modules would look like in order to be more efficient, responsive and interlinked. The objective of the paper is to launch a reflexion on what would be the implications for the data collection systems if such survey architecture would be adopted. Therefore, it concludes with an initial reflexion on what would be the needs of an integrated survey design from the data collection system, and also what opportunities would be open.

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

II. The industrialisation of statistical production

4. Official statistics face challenging conditions nowadays. Users want more and faster statistics, while official statistics producers face decreasing funding but also new opportunities made possible by technological advances. The UNECE High-Level Group for the Modernisation of Statistical Production and Services (HLG) has identified, in its strategic vision[2], the industrialisation of statistical production as the answer to those conditions.

5. Standardisation is the key to this industrialisation process. Standard statistical processes, tools and methodologies are developed only once and re-usable from one statistical domain to the other (UNECE, 2011). The decrease in diversity of these elements and increase in quantity of standardised elements increases the potential gains of automation in the production of statistical products, leading towards a separation of the processes of production and design, with the subject matter expertise focused in the latter (Kent, 2011).

6. Shortly before the Conference of European Statisticians (CES) endorsed the strategic vision of the HLG in June 2011, Eurostat had put forward a vision for the next decade on the production method of EU statistics in the form of a Communication from the Commission to the European Parliament and the Council[3]. It called for the modernisation of the production of European statistics based on the replacement of the “augmented stove-pipe model” by an “integrated model”.

7. Under the “augmented stove-pipe model”, the several national statistical products are produced on independent production lines generally separated by statistical domains and European statistics are based on the compilation of national statistics, similarly on a domain by domain basis. In contrast, in an “integrated model” the statistical products in the several statistical domains would be produced as the result of a single comprehensive production system, at both national and European levels.

8. The “vision”, as it became known in the European Statistical System, is Eurostat’s answer to the changes in the environment of statistical production in Europe. Firstly, users need not only more and higher quality statistics, but also more interlinked statistics which allow the proper assessment of phenomena which are transversal to statistics domains. Secondly, the reduction of administrative burden, including statistical burden of survey respondents, is a goal of the EU. Thirdly, in the last decades we have witnessed significant developments in the technology for dealing with information, including statistics, which enlarge the possibilities space for statistical products and allow significant efficiency gains in information capture, processing and dissemination.

9. In social statistics the “vision” was translated into a strategy via a Eurostat and NSI joint approach to modernising European Social Statistics. The adoption of the strategy was formalised in the Wiesbaden memorandum[4] agreed by the Directors-General of Eurostat and of the NSI of the EU Member-States in September 2011. The strategy has as key actions to promote reliable and up-to-date sampling frames of individuals and dwellings, better access to administrative data, to streamline the social surveys around a limited number of pillars and to increase the use of new technologies and statistical tools. This paper focus on the experience so far with the action which has particular relevance for data collection, the streamlining and integration of the European social surveys.

III. Survey integration

10. In this context of industrialisation of statistical production and of modernisation of European social statistics, the streamlining and integration of the European social surveys has three objectives: 1) increase efficiency; 2) increase responsiveness to users’ needs; and 3) increase the level of integration of the system of social surveys. The level of integration is to be understood here as the extent to which information from different statistical domains can be crossed.

[2] http://www1.unece.org/stat/platform/display/hlgbas/Strategic+vision+of+the+HLG
[3] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF
[4] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/de/DGINS2011_memorandum.pdf

11. There are eight European social surveys[5] covering mostly the labour market, the income, social inclusion and living conditions, the education, the health and the information society statistical domains. They consist of micro-dataset specifications, i.e. list of variables to be obtained for each statistical unit of a sample of a reference population. They are accompanied by precision requirements and a set of methodological conditions to be followed by the sampling design or the survey vehicle used at national level to collect the data. The data sources at national level can be surveys or registers with information at individual level. Currently the main sources are surveys. The European social surveys differ from some other European statistics specifications, as they define intermediate statistical products, micro-datasets, which are then used at European level to produce the final products: indicators, tabulations and statistical analysis.

A. Modularisation of the European social surveys

12. The idea of modularity was introduced, even if only in a mitigated form, by the expert group set up by the DSS (group of directors of social statistics of the European Statistical System) to investigate the streamlining of the European social surveys (Zhang, 2011). While trying to re-think the composition of the eight European social surveys and recomposing them into a smaller number of “pillars”, it became evident that the surveys per se were too large to serve as building blocks. Therefore, it was concluded that a deconstruction of the content of the current surveys into modules was necessary. A reconstruction of the European social surveys would then be possible based on those modules. A process of modularisation, i.e. decomposition of the content of the surveys in terms of micro-data variables into modules, was then initiated.

1. Modularisation criteria

13. The immediate idea was to define a set of criteria which would guide the definition of the modules. The criteria reflected centrifugal forces, those that prevent content from being part of the same module, and centripetal forces, those which require variables to be kept in the same module (Eurostat, 2012a). The expert group selected 13 modularisation criteria (Table 1), including both inherent characteristics of the variables, as the statistical unit to which the variable refers, as well as methodological characteristics, such as the sample size or the mode of collection.

Table 1. Modularisation criteria selected by expert group on streamlining and integration of the European social surveys

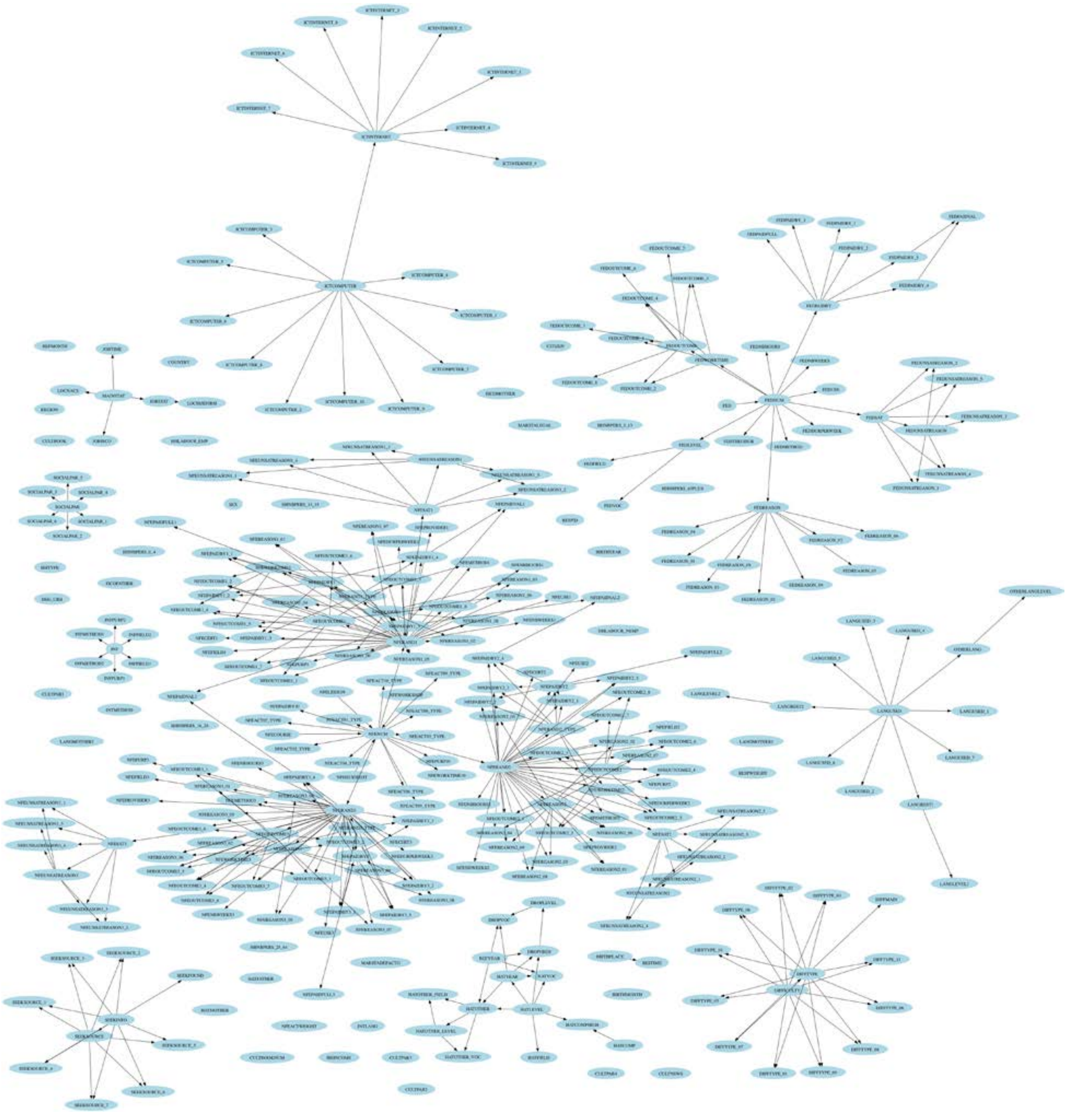
Topic
Statistical unit
Household clustering
Context (validation and answering aid)
Accuracy/Sample size requirements
Reference period
Data collection period
Mode of collection
Frequency
Target population (including filtering)
Need for a panel / longitudinal component
Other restrictions in terms of methodological parameters
Crossing needs

Note: Check annex for detailed explanation of each criterion.

[5] The eight European social surveys are: Labour Force Survey (LFS), Statistics on Income and Living Conditions (SILC), Adult Education Survey (AES), European Health Interview Survey (EHIS), European Health and Social Integration survey (EHSIS), Community Survey on Information and Communication Technologies Usage in Households and by Individuals (ICT), Household Budget Survey (HBS) and Harmonised European Time Use Survey (HETUS).

14. The first attempt to use the modularisation criteria to group micro-data variables in modules (i.e. split the complete list of variables of the survey in modules) was done in education statistics, with the Adult Education Survey (Eurostat, 2012b). The main lesson taken from that first attempt was that the application of the criteria was highly conditioned by the filtering and routing of the variables (included in the target population modularisation criterion). The sequence of filters between variables led to the grouping of the variables with their corresponding filters forming "natural" modules (Figure 1). Although the filtering did not fix completely the composition of the modules, it significantly conditioned the use of the other modularisation criteria.

Figure 1. Network graph linking AES variables to their filter variables



15. Even if the filtering in the AES conditioned the use of the modularisation criteria, at least it allowed the creation of separate modules without dependencies[6] between each other, a desirable characteristic if modules are to be re-used from one survey to the other (more on this in section II.B). However, the same exercise was done for the ICT survey and in that case a set of modules without dependencies

[6] Dependency between modules here means that the use of one module requires the use of another because a variable in the first is filtered by a variable in the latter.

could not be defined. All the variables in the ICT survey are filtered by some other variable in cascade, tracing back to two basic variables: computer availability and internet access. This experience has shown that, firstly, it is most likely that it is not possible to reduce the European social surveys to a set of statistical modules without dependencies between them and, secondly, that the modularisation involves not only the grouping of existing variables but also their revision, in particular a re-think of the way we use filtering and routing in our surveys. That is what was done in the case of the LFS. As the LFS is coincidently under revision at the time of writing, its modularisation includes a revision of the filtering and of some of the variables (Van der Valk, 2012a, 2012b). Even so, although the filtering was simplified, it was not feasible to define the modules of the LFS without any dependencies between them.

Figure 2. Education attainment module designed after first attempt to modularise the AES

Variable code	Name	Filter
HATLEVEL	Highest level of education or training successfully completed	
HATVOC	Orientation of the highest level of education or training successfully completed	HATLEVEL=22 to 40 and (REFYEAR- HATYEAR) ≤ 20
HATFIELD	Field of the highest level of education or training successfully completed	HATLEVEL=22 to 60
HATYEAR	Year when highest level of educaton or training was successfully completed	HATLEVEL<>01, -1
HATOTHER	Other formal education or training successfully completed in another field than 'hatlevel'	HATLEVEL=22 to 60 and (REFYEAR- HATYEAR) ≤ 20
HATOTHER_LEVEL	Level of the formal education programme	HATOTHER=1
HATOTHER_VOC	Orientation of the formal education programme	HATOTHER=1 and HATOTHER_LEVEL=22 to 40
HATOTHER_FIELD	Field of the formal education programme	HATOTHER=1 and HATOTHER_LEVEL=22 to 60
HATCOMP	Procedure of recognition of skills and competences undertaken	
HATCOMPHIGH	Recognition of skills and competences allows access to a higher formal education programme than the level mentioned in 'hatlevel'	HATCOMP=1,2 and HATLEVEL<>01, -1
DROPHIGH	Formal education abandoned higher than the level mentioned in 'hatlevel' but not completed	HATLEVEL<>01, -1 and (REFYEAR-HATYEAR) ≤ 20
DROPLEVEL	Level of the formal education not completed	DROPHIGH=1
DROPVOC	Orientation of the formal education not completed	DROPLEVEL=22 to 40 and (REFYEAR- HATYEAR) ≤ 20

Note: Darker rows refer to derived variables

2. Standardisation of variables and modules

16. Central to the idea of streamlining the social surveys around a limited number of pillars, was that efficiency gains could be obtained by avoiding the collection of the same variables, or of variables not exactly the same but with only slight differences, in several surveys. The removal of some of those "repeated" variables could be used to introduce new variables answering to new users' needs (fulfilling the objective of increasing the responsiveness of the system of surveys), could be used to reduce respondents burden, or both. However, that is not possible without hampering the goal of having a system of surveys allowing crossing information between statistical domains. Therefore, that has to be done by a critical review of the existing joint distributions (of micro-data variables) which are most important and should therefore be kept in the surveys, and of harmonisation of those variables which being kept in several surveys had some small differences which could be eliminated (sometimes similar variables in different surveys follow different definitions for very good reasons).

17. The modularisation has to take into account this need of having variables used in more than one survey. Therefore, modules have to be created taking into account if their variables are used in several surveys or if they are specific to a particular survey. The solution is to create standard modules mostly composed of variables used in several surveys. The modularisation is therefore accompanied by a process of standardisation. The creation of standard modules requires a re-think, once again, of the way we design the content, i.e. variables and modules, of our surveys. Standard variables and modules need to be generic enough so they can be introduced unchanged in several surveys and still integrate properly with the rest of the survey content.

3. Modularisation: a work in progress

18. Eurostat is currently working on the development of proposals of standard modules for those variables of widespread use in the European social surveys. The starting point is the core social variables, background variables which are mandatorily included in every European social survey. They are mostly demographic variables, but include also education level, labour status and income. A module under discussion is on the demographic background (Figure 3) which includes all demographic core social variables[7]. The work on this module has taught us another lesson. While for domain specific modules the filtering between variables guided the design of the module, in this case there is hardly any filtering involved. The construction of the module was based on the topic (demography) and on how frequently it is used in the surveys.

Figure 3. Standard demographic module under discussion

Code	Name	Filter	Core	LFS	SILC	AES	EHIS	EHIS	HBS	HETUS	ICT
SEX	Sex		X	X	X	X	X	X	X	X	X
BIRTHDATE	Date of birth (DDMMYYYY)										
BIRTHYEAR	Year of birth			X	X	X					
BIRTHMONTH	Month of birth				X	X					
BIRTHDAY	Day of birth										
AGE	Age		X	X			X	X	X	X	X
BIRTHPLACE	Country of birth		X	X	X	X	X	X	X		X
ARRIVYEAR	Year of arrival in the country	BIRTHPLACE <> 00			X						
YEARESID	Years of residence in the country			X		X					
CITIZEN	Country of main citizenship		X	X	X	X	X	X	X		X

Note: Darker rows refer to derived variables

19. The inclusion of 'year of arrival in the country' in the standard demographic module raised an important question. Do the modules have to be used in the surveys without any change? Ideally, modules should be taken exactly as they are defined in the standard, so consistency between surveys is kept. However, the variable 'year of arrival in the country' is the only variable which is not a core social variable and including it in a module composed almost completely of such variables would imply that it would be included in all surveys, leading to an unnecessary increase of burden. Taking it out of the module would result in having a new module composed of a single variable, as there are no other demographic variable, and it would be a module dependent on the demographic module because 'year of arrival in the country' is filtered by the variable 'country of birth' which would be kept in the demographic module. Therefore, the best solution seems to be keeping 'year of arrival' in the demographic module allowing each survey to adapt the module by excluding it if there is no need to collect the variable.

20. The key question for those of us working on the modularisation of surveys is what a statistical module really is. The work so far in Eurostat shows us that it is still too soon to answer definitely to such question. There are some lessons learned, such as the key role of filtering, the need to allow for dependencies between modules and the need to allow some flexibility on the adaptation of the modules. However, a definitive answer will also depend on how the modules will be combined into surveys and what their role in the several steps of statistical production will be.

B. A modular survey architecture

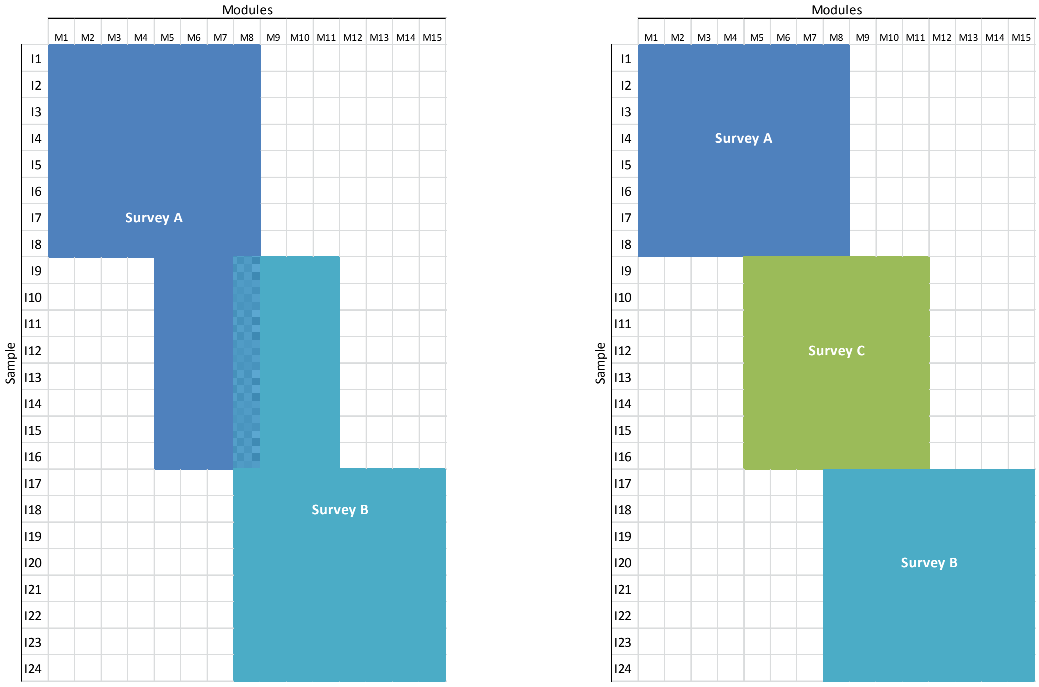
21. The expert group on the streamlining and integration of the European social surveys not only identified the modularisation as a necessary step in the re-organisation of the surveys, but it also recognised that their integration would raise operational and methodological issues. The main operational issues identified were the impact on case management systems and the management of the field force. The methodological issues raised concerned the impact on sampling design and on estimation. In order to gain a better understanding of the implications of an integrated system of surveys on sampling design and estimation, Eurostat started a project with the objective of assessing past experiences of survey integration, review research literature and develop methods and scenarios of implementation for an integrated system of European social surveys.

[7] Except household composition and marital status which should be included in a household composition module.

22. The initial work of the project revealed that there are three alternative ways to deal with sampling design and estimation in an integrated system of surveys (Eurostat, 2013a). Firstly, the information needs of all statistical domains could be combined in one single multi-purpose survey (one single sampling design and one estimation procedure). Secondly, the surveys could be kept separated with their own sampling design, but with partially overlapped samples (Figure 4a). Thirdly, surveys can be kept separated, with separated sampling designs and separated sample selection, leaving integration, and efficiency gains, to the estimation phase. The third approach seems to be the most appropriate for an overall system of social surveys. Leaving the integration to the estimation phase, by combining the data collected from the several surveys, has the advantage that the sampling can be optimised to the specific data to be collected, avoid the long questionnaires covering many disparate subjects of multi-purpose surveys and keep the decision in terms of sample design and survey content separated which would not happen with partially overlapping samples (Eurostat, 2013a).

23. It is important then to clarify what delaying the integration to the estimation phase really means. The pursuit of efficiency gains in an integrated system of surveys can be seen as a problem of finding and eliminating excessive sampling, i.e. the collection of data on a variable from a sample which is bigger than it needs to be to obtain results with the desired level of precision. That excessive sampling exists today in our survey systems from basically two sources. Firstly, we collect some variables simultaneously in several surveys. The sample size of each of the surveys is determined in order to provide accurate estimates for the variable. Therefore, in total we collect the data from double the number of individuals which is actually necessary. Secondly, sample sizes are determined to provide accurate results for all the variables in the survey[8]. However, every variable has its particular needs in terms of sample size, either because of statistical properties of the variable, or because of particular results we want to estimate from the variable. Therefore, for all variables, apart from the most demanding one, data is collected from a number of individuals higher than what is actually necessary.

Figure 4 – Integrating surveys



(a) More efficiency with partial sample overlap

(b) More efficiency with combination of results from separate surveys

24. How can we recover that excessive sampling in the context of an integrated system of surveys based on the combination of data from separate surveys? The answer can be illustrated by Figure 4. Some modules are collected in more than one survey (M8 in figure 4a) providing accurate estimates in each one of them, while only one would suffice. The modularisation should enable us to identify modules

[8] Sometimes in practice this is done by determining sample sizes based on a reference variable known to be particularly demanding in terms of sample size.

for which samples can be smaller while still allowing us to draw accurate estimates (modules M1 to M4 and M12 to M15 in figure 4.a). Some sampled individuals would then receive smaller questionnaires (e.g. individuals I9 to I16). Those individuals could be included in the sample of another survey (survey B in Figure 4a) and the data on module M8 collected in survey A could then be re-used in survey B. Additionally, new crossings between modules M5 to M7 from survey A and modules M9 to M11 would be possible. The survey organisation in Figure 4a can be reproduced as the combination of 3 separate surveys (Figure 4b). Instead of having a long and a short questionnaire, survey A would have only the long questionnaire and cover only the smaller sample needed by modules M1 to M4. The same would happen to survey B, including only a long questionnaire and covering only the sample required by modules M12 to M15. In order to be able to produce accurate estimates for modules M5 to M11, surveys A and B would be complemented with survey C. The new survey C would play a double role. Not only it provides the sample bust required by modules M5 to M11, but it also provides crossings between those modules. This illustration shows that an architecture of social surveys based on standard harmonised modules across the system and the combination of datasets can deliver more efficiency and more crossings between modules.

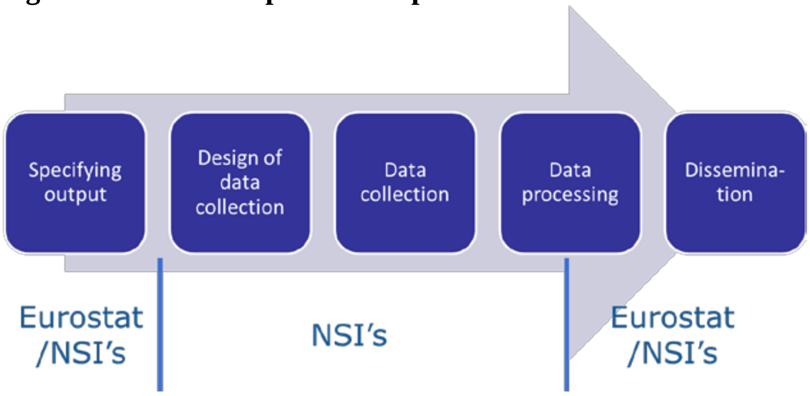
25. The understanding of how an integrated system of surveys can be more efficient, providing at the same time more crossings, sheds some light on how such a system, based on modules, would look like. A modular system of surveys would be composed of modules, i.e. groups of variables at micro-data level, which are combined into instruments providing the crossings which are necessary to produce the intended results (Eurostat, 2013b). Each instrument would then be collected from its specific sample and results for a particular module would be estimated using the data collected in all instruments which use that module. Methods for the optimal composition of instruments, determination of their sample sizes and estimation of results, for such a system of surveys, are being investigated at Eurostat.

26. Such a modular architecture would be a “plug-and-play” system where standardised statistical modules would be developed once and used to compose statistical products (the instruments) answering to specific statistical needs[9]. Having this architecture in mind, we get a better understanding of what a statistical module is. Modules are to be re-used from one survey to the other and therefore need to be generic enough so they can be used with minor adaptations. Data for the same module from different instruments would be combined. Therefore, the impact on the results of including the modules in different instruments needs to be minimised.

C. Modularity in the several stages of statistical production

27. The discussion about statistical modules and a modular system of surveys, in particular linked to an efficient, flexible and crossings richer system, has focused particularly on the specification of outputs at European level. However, the set-up of a modular system of surveys has implications also at national level and in all steps of statistical production (Figure 5) (Van der Valk, 2013).

Figure 5 – Statistical production process



Source: Van der Valk (2013)

[9] In such architecture, surveys can be seen as groups of instruments sharing the same subject matter and the same methodological characteristics.

1. **Specifying output**

28. In a modular survey architecture, outputs would be specified in a finer grained level. Instead of being primarily linked to complete surveys, indicators would be linked to modules or to a restricted number of modules (Van der Valk, 2013). Tabulations would be linked to required joint distributions, modules that need to be collected together from a sample of individuals. Precision, frequency and longitudinal requirements of indicators and tabulations would then be translated into requirements of individual modules and of combinations of modules. Output specifications would then result on micro-datasets specifications composed of modules, with modules being present potentially in several micro-datasets. These output specifications would happen at European level but also at national level where national specific needs could result in higher precision requirements for modules and combinations of modules or on new joint distributions between European harmonised modules or with new national modules.

2. **Design of data collection**

29. When data collections are designed national level, European output specifications and national output specifications would then be combined. At this stage new variables needed for specific national needs would be incorporated in European standard modules or in new national specific modules. Modules would still be the building blocks to define the national instruments which would provide the data required for European and national needs. Samples or subsamples would be drawn for each instrument. The mode of collection or statistical source could be adapted to the type of data to be collected, instead of complete surveys, by assigning the mode of collection to particular modules (Van der Valk, 2013). For example, a module on income could be obtained via tax registers, instead of being incorporated in survey questionnaires. Once a mode of collection or source is optimized for a particular module that mode, or source, would be used for whatever instrument which uses the module.

3. **Data collection**

30. Data would not be collected module by module but by instrument[10]. However, the architecture would allow an easier design of more targeted statistical products (instruments). Therefore, it would normally result in a much higher number of different instruments and corresponding questionnaires. Case management systems and fieldwork would need to be prepared to handle the potentially much higher number of different questionnaires. However, the modular design of the instruments helps the data collection systems handle such increase in diversity of questionnaires. With harmonized modules between instruments electronic questionnaires can be designed also in a modular way, with sets of questions for a module being designed only once and reused from one questionnaire to the other. Manuals and other interviewers training material could also be produced only once per module. Well defined standard modules would also be the building blocks of data collection.

4. **Data processing**

31. Validation rules and methods can be refined to the particular data of a module and developed only once and run whenever the module is present in an instrument. The design of the modules should take into account the validation needs of their variables ('context' modularisation criterion). So, when some variables are used to validate each other, they should ideally be kept in the same module. This should cover most validation needs. However, there will always be cases where variables in different modules can be used to validate each other. In those cases validation rules should be setup and triggered whenever the modules are present in the instrument. The use of standard harmonised modules would allow the automation of such triggering. For modules which data would be collected from administrative records, the set of procedures to validate and align the administrative data to the statistical concepts would be developed once and shared between different instruments and statistical domains. Finally, data processing based on the combination of datasets would require a data warehouse approach from some stage of the process.

[10] In the case of modules which source are administrative records, they could be treated module by module.

5. Dissemination

32. Concerning what is disseminated, results would be more harmonised between different surveys and statistical domains because standard modules would be used and grand totals would be aligned to one single figure. There would be more opportunities for disseminating results as more crossings would be available. Concerning the process of dissemination, considering that questionnaires would be collected at different moments in time and at different speeds, results at a certain moment in time would depend on what datasets are available. This would require a dissemination strategy with leading results based on the first datasets to be available (for some statistics possibly sooner than is the case nowadays) and a revision policy based on the availability of additional datasets.

IV. Centralisation of data collection

A. Needs

33. A modular survey architecture will pose particular needs to the data collection systems. In order to attain a more efficient, more flexible, more responsive and more interlinked survey system we will need to deal with a potentially much higher number of different questionnaires. Questionnaires would be designed to target needed joint distributions. Only electronic modes of collection will be compatible with such diversity. Managing a significantly higher number of paper questionnaires would pose costs too high. Case management systems allocating sampled units to interviewers would need also to be prepared to deal with the increase of the number of different questionnaires.

34. For each questionnaire samples will be smaller. This might have implications for the management of the field force when personal interviewing is used. For example, it might not be efficient anymore to have interviewers specialised in particular questionnaires because in the geographic zone he or she covers very few persons would be allocated for each questionnaire. On the other hand, questionnaires would be composed of harmonised modules, so interviewers could specialise in a large set of modules.

35. Data collection would not serve particular surveys but would instead input a large more complex data processing system. The interoperability between the IT systems supporting data collection and those supporting data processing would be fundamental. Data processing would be dependent on the data collected via different questionnaires and the quick transfer of data would be a key feature, otherwise results will be as timely as the slowest questionnaire. The implications of data collection delays on the dissemination of results would also be less evident because each questionnaire would be used in the production of several statistical products.

36. The approach of data collection will be less that of collecting data for a certain survey, and more of maximizing the 30 minutes, or whatever is the level of tolerance of respondents, to collect the "right" data. The focus would not be to collect the maximum amount of information from the respondent because for some variables enough information has already been collected. The focus would be to assign the proper questionnaire (instrument) to the individual.

37. The data collection system needs to be able to assign the most appropriate mode of collection not only to the phase of contact with the respondent but also to the module being collected (because of the type of data in the module, particular modes of collection could be assigned to it). This means that a single respondent could be contacted via different modes depending on the module. For example, an initial personal interview could collect information requiring only memory and for data for which the respondent need to consult some records (e.g. expenditures) a web questionnaire could be used.

38. The data collection system would need to be centralized in order to provide an answer to these needs. A decentralized system would not be able to serve a very high number of questionnaires, with smaller samples and fast enough for such integrated system of surveys.

B. Opportunities

39. A modular system of surveys also offers opportunities to data collection systems. While the number of different questionnaires would increase and the sample size to which they would be applied would decrease, questionnaires would be more targeted to the joint distributions which are really needed. There would be no need to stack up variables in a survey questionnaire because that is the vehicle available. Questionnaires could be shorter providing an answer to the decreasing response rates and decreasing respondents' tolerance to long interviews.

40. The use of standard modules would increase the level of harmonisation and this could lead to an increase of the efficiency of the data collection system and quality of the data collected. It would mean fewer variables which are very similar but still different between surveys. It would also mean fewer opportunities for confusion for the interviewers and training would be simpler.

41. The modular nature of the survey designs would be easily transposed to questionnaire design. Questions could be designed by module and questionnaires composed by putting together those questions. Questionnaire design would also be modular and that would make it cheaper and easier to deal with a large number of different questionnaires.

42. Also manuals, explanatory notes and other supporting and training material could be organized by module. Once created for a module it could be reused for all questionnaires which includes that module.

V. References

Eurostat (2012a). Results of the expert group consultation on the criteria for the modularisation of the European social surveys. Second meeting of the Expert Group on the Integration of the European social surveys

Eurostat (2012b). First steps in the modularisation of social surveys – the case of the Adult Education Survey. Second meeting of the Expert Group on the Integration of the European social surveys

Eurostat (2013a). Implications of the methodological pilot studies on survey integration for the development of a modern architecture of European social surveys. Third meeting of the Expert Group on the Integration of the European social surveys

Eurostat (2013b). Methodological approach to a modules based system of European social surveys. Third meeting of the Expert Group on the Integration of the European social surveys

Van der Valk, J. (2012a). Introducing modularity in order to improve quality and efficiency. 7th Workshop on LFS methodology, Madrid, Spain, 10-11 May 2012.

Van der Valk, J. (2012b). Improving the quality of complex surveys: The case of the EU Labour Force Survey. European Conference on Quality in Official Statistics, Athens, Greece, 30 May-1 June 2012.

Van der Valk, J. (2013). Modernisation of Social Statistics - A possible future of the European social surveys. *Mimeo*

Zhang, L. (2011). Towards an integrated module-based social survey design. *Mimeo*

Annex – Modularisation criteria (explanatory notes)

Topic

The thematic grouping of variables is done basically for two reasons. Firstly, variables on the same subject matter are analytically related and therefore should be collected together. Secondly, collecting variables thematically related facilitates the answering by the respondents.

Statistical unit

In the context of the modularisation the statistical unit (individual or household) has a role similar to that of the topic. Separating variables referring to households from those referring to individuals facilitates the answering and avoids reporting to the wrong unit.

Household clustering

For some variables there are reasons to collect them for all members of the household (e.g. income and consumption). Household clustering (i.e. collecting variables from all members of the household) has a negative effect in the efficiency of the sampling. Therefore, variables with such restriction should be separated from other variables, so that the sampling inefficiency is not passed to variables unnecessarily.

Context (validation and answering aid)

Two issues are related to the context of the variables:

When variables are collected via surveys (instead of for example obtained from administrative records) the response depends many times on with which other variables it is collected. Different variables (i.e. questions) imply different memory recall processes and what may not be recalled with one question can be recalled by another related but different question.

ii) The validation (checks) between variables also depends on what variables are collected together. If the checking possibilities are not taken into account, the modularisation can lead unintentionally to a decrease of the accuracy of the data collected.□Context has implications at two levels. Firstly, it is relevant for the definition of the modules. They should group variables which provide appropriate and sufficient context to each other. Secondly, it is relevant when deciding what modules should be combined. In the latter case, the interaction between modules can be positive or negative. It would be negative in those cases where some variables mislead the understanding of some other.

Accuracy/Sample size requirements

The grouping of a long list of variables collected from one single sample implies that all the information is collected from as many units as it is required by the most demanding variable. Therefore, variables with significant different sample size requirements should be whenever possible separated into different modules to which different sample sizes can be applied.

Reference period

Questionnaires should cluster variables with the same reference period. Putting close together questions which refer to different reference periods might create confusion for the respondents and the information collected not refer to the period which was intended for each variable. This criterion is less relevant when the information is not collected via questionnaire, but via administrative records.

Data collection period

Variables with particular requirements concerning the data collection period should be identified. This will be very important in the programming phase which will have to be set up, after an architecture of social surveys is agreed.

Mode of collection

The mode of collection has a significant impact on the data collection cost. When the mode of collection is restricted to a more expensive mode (e.g. personal interview instead of telephone) for some variables, it is important that those variables are clustered into a module leaving other variables free from such restriction.

Frequency

Variables which are needed with different frequencies should be grouped separately. The reduction of the frequency to what is strictly needed has a very significant impact on the cost, in particular when the data is collected via questionnaire.

Target population

The target population includes both what are currently the target populations of the different surveys and also the filtering of some variables by the values of others. Filtering is a very important tool to limit the burden of the respondent. However, it poses important restrictions to how the components will be combined. Namely, components which are filtered need to be combined with the component which includes the filter variable. The filters also partly define the flow of the questions in a questionnaire, which might also need to be taken into account.

Need for a panel / longitudinal component

For some variables it is particularly important to have repeated measures over time for the same sampled unit. This is the case when the persistency of a phenomenon (e.g. poverty and unemployment) is particularly relevant. When this is the case a panel needs to be used and a particular sample could be used for these variables, probably different from the one used for variables which only require a cross-sectional sample. The maintenance of a panel sample is particularly costly. Therefore variables which do not require a panel should be separated whenever possible from those which require it.

Other restrictions in terms of methodological parameters

There are other restrictions or characteristics of variables which condition how they can be grouped. That's the case of, for example, a specifically required timeliness, the impossibility of using proxy interviews and restrictions in terms of mode of collection.

Crossing needs

Variables belonging to the same component will be collected always together, i.e. for the same individuals. They will be collected together with variables from different components only when their correspondent components are combined. All the variables of a certain component will share the same list of variables with which they will be collected together. Therefore, variables should be grouped also based on with which variables we want them to be crossed.