UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

Editing process and its quality regarding design and production phases using process metadata and calculation modules

Prepared by Pauli Ollila, Statistics Finland, Finland

I. Introduction

1. Some main aspects of developing and implementing the editing process are studied by using concepts defined for different levels of the editing process. The development in the field of editing at Statistics Finland is presented within that framework, emphasizing the experiences on evaluation and improvement.

2. Chapter II provides a study of four phases of editing strategy development, regarding designing, IT realization, testing and production. The locations of these phases are demonstrated in Generic Statistical Business Process Model to illustrate various ways when developing the editing strategy. The work on the structure of editing process has been carried out in a UNECE task team. This framework is studied in terms of phases of editing strategy development, emphasizing different needs of designing, making realizations, testing and carrying out production. Finally, we study editing in production following the process flow and process steps and connections to suitable methods with parameterization and proper IT solutions. The role of metadata in different forms is also taken into account.

3. Chapter III takes a look at monitoring and evaluation in general, either in production or test situations. The process metadata system forms the basis of measuring the properties and the quality for monitoring and evaluation. Also non-measurable evaluation is dealt with in this context. In the last part of Chapter III we study the parts in the process to which the improvement can be targeted. This seven-class categorization is based on process levels appearing in the task team work and the IT system realizations of these levels. Examples of each class are presented.

4. Chapter IV presents implementations and experiences in editing processes at Statistics Finland at various levels with examples, following the framework to study evaluation and improvements provided in Chapter III. The paper includes references to material and definitions from *"Task Team on a Generic Process Framework for Statistical Data Editing - Generic Statistical Data Editing Models (GSDEM), draft prepared for the 2015 UNECE Work Session on Statistical Data Editing"*, marked (GSDEM 2015).

II. Development and production in editing

A. General view at developing and carrying out the editing process

5. The *editing process* "contains a number of activities or tasks that aim to assess the plausibility of the data, identify potential problems and perform certain selected actions that intend to remedy the identified problems" (GSDEM 2015).

6. The *editing strategy* covers in addition to the editing process also many other things connected to the operations dealing with the editing process in different contexts, and thus it can be considered to be a wider concept than the editing process.

7. Here we recognize four different *phases of editing strategy development*. These phases are *designing the editing strategy, Creating the realization system of the strategy, Testing the strategy* and *Applying the strategy in production*.

8. *"Designing the editing strategy"* includes a variety of different planning and decision actions for editing, ranging from the construction of the process flow to selecting some principles to carry out edit rules. Although the term "strategy" may be too grand to describe editing in some statistics, the editing principles and realizations still form a strategy.

9. *"Creating the realization system of the strategy"* includes the IT choices for carrying out the editing strategy, both containing solutions for different parts in the editing strategy and possibly decisions for interactions between different IT environments and data structures. In principle this phase could also include decisions about non-IT practices belonging to the strategy, e.g. paper questionnaire studies.

10. *"Testing the strategy"* includes operations in the realization (IT) system of the editing strategy with specific test data sets (unedited, edited), in practice mostly from earlier rounds of the production of statistics. The testing procedure may have a systematic structure, but known statistic-dependent problematic areas can also be studied. This phase should help in both evaluating and improving the editing process and the parts of the IT system.

11. *"Applying the strategy in production"* contains the implementation of the editing strategy with chosen methods and parameterization in the corresponding IT system with data set(s) to be edited in the production process of statistics. Note that editing in production can also happen in a collection phase, i.e. when we are acquiring the data gradually.

12. These four phases can be found in different places in the process of statistical production. How can they be located to a process model framework for statistics? In Figure 1 the phases of editing strategy development are highlighted in the pattern describing the *Generic Statistical Business Process Model*, i.e. GSBPM (Vale 2011).

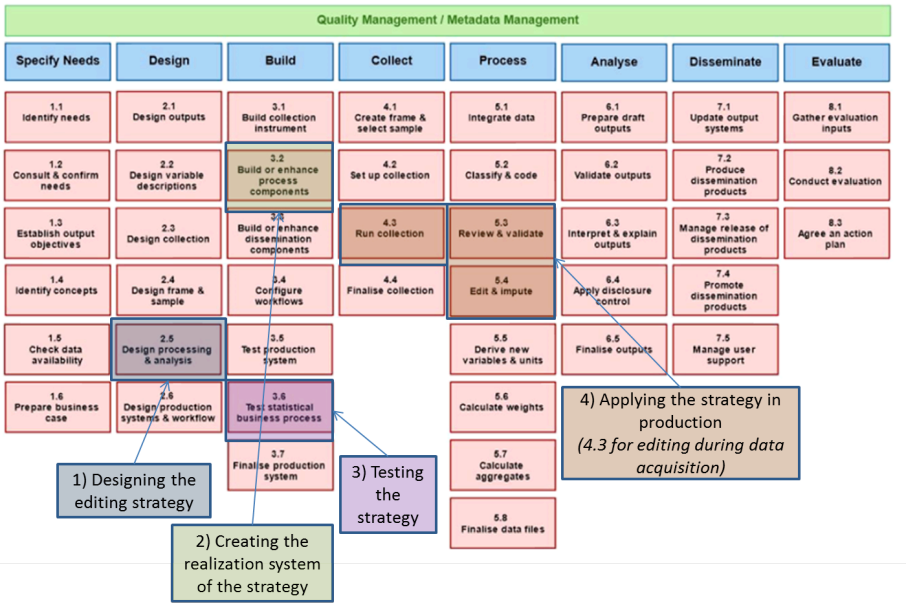

Figure 1. Phases of editing strategy development in GSBPM

B. Different levels of editing process and phases of editing strategy development

13. The editing process and its different elements have been of interest in recent years (e.g. Saint-Pierre and Bricault 2011, Pannekoek and Zhang 2012, Ollila et. al 2012). The task team work to be

presented in 2015 UNECE Work Session (GSDEM 2015) concentrates on these issues. In this chapter this framework is studied in terms of the phases of editing strategy development, presented in the previous chapter, emphasizing the different needs of designing, making realizations, testing and carrying out production. Several GSDEM definitions of editing terms are used in this chapter with reference.

14. The *overall data editing process* contains a number of *activities* or *tasks* that aim to assess the plausibility of the data, identify potential problems and perform certain selected actions that intend to remedy the identified problems. The process itself can be structured by splitting it up into sub-processes, called *process steps*, and a *process flow* that describes the navigation among the process steps during execution. (GSDEM 2015).

15. An operational data editing process usually contains a considerable number of functions with specified methods that are executed in an organised way. In terms of purpose, the different *functions* involved in editing can be categorised into three broad categories:

(a) *Review*. Functions that examine the data to try and identify potential problems. This may be by evaluating formally specified quality measures or edit rules or by assessing the plausibility of the data in a less formal sense, for instance by using graphical displays;

(b) *Selection*. Functions that select units or fields within units that may need to be adjusted or imputed or, more generally, identify selected units or variables for specified further treatment;

(c) *Amendment*. Functions that actually change selected data values in a way that is considered appropriate to improve the data quality. This includes changing a missing value to an actual value, i.e. imputation.

A *function* is an instance of one of the three function types that serves a specific purpose in the chain of activities that leads to the edited data. (GSDEM 2015). The idea of three categories of functions can be found also e.g. in Pannekoek and Zhang (2012) and Pannekoek et al. (2013) with slightly different terminology. The process steps defined for the process include one or more functions each.

16. Data editing functions specify *what* action is to be performed in terms of its purpose, but not *how* it is performed. The latter is specified by the *process method*. Some methods can perform different functions at once. (GSDEM 2015). For the purposes of production some computerized methods are *parameterized* in order to define the process exactly. Methods can be interactive as well, and then there is no parameterization as such. Note that the parameterization can sometimes describe the sequence and choices in a process step or in the process flow.

17. The *planning* of the process with different levels of actions defined does not tell how to carry out the process in production. Further, the act of *deciding* the methods (based on our knowledge on methods) with ideas of parameterization is needed. These two entities (*planning* and *deciding*) form the phase *"Designing the editing strategy"*.

18. However, these are not sufficient for production without *developing an IT system* to enable realization of the decided methods and parameterization (phase "*Creating the realization system of the strategy")*. There might be various ways to carry out the methods, but the corresponding IT solutions for the methods are denoted here as *modules*. For interactive actions there won't be modules as such, but the IT environment could provide *modules or solutions to support the interactivity (e.g. applications)*.

19. The IT system with decided methods and parameterization must be tested with some data sets available in order to have the system ready for the production (phase "*Testing the strategy*"). These tests can alter either methodological definitions or IT solutions or both. Finally, the *production phase* is a string of actions carried out in the suitable system with defined methods and parameters following the process flow with process steps and functions addressed to tasks in the process steps (phase *"Applying the strategy in production"*).

20. Figure 2 illustrates the situation with terms needed to be decided in each phase (order not included in the process flow and process step parts).

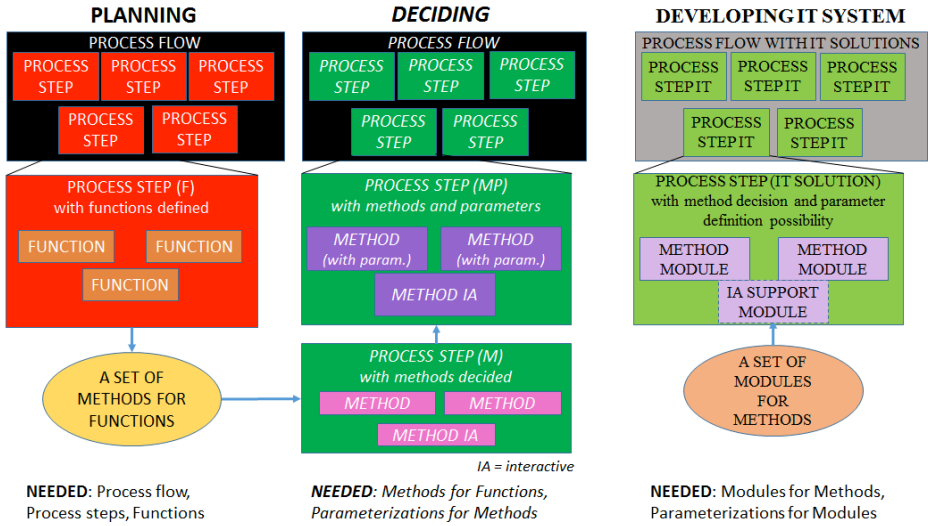

Figure 2. Different levels of editing process in terms of planning, deciding and developing IT system

21. The vision of parallel-type of planning, deciding and developing the IT system presented in Figure 2 is not met in practice in most cases. The process flow can be developed during production years with gradual improvements and as reactions to various requirements. The developmental efforts might be reactions to some problematic situations or adjustments to a more general process rearrangements or changes in the IT environment. There is not necessary enough knowledge about existing methods and good practices, and the IT environment does not provide suitable modules or the construction of the modules is too laborious. Instead, some temporary-type and non-parameterized solutions with programming efforts needed from one round to another might be still selected.

C. Carrying out the process flow of editing in production

22. The process flow and its process steps with specified functions are outlined in the planning phase. The methods chosen in the decision phase are supposed to be suitable for the functions. The preliminary parameterization of methods and process steps can be decided in that phase as well. The IT system solutions provide the possibility to carry out this entity for *editing in production, targeted to the statistical data*.

23. In GSDEM (2015) the *statistical data* is defined as "the data that is the object of the editing activities". When a step can be followed by several alternative steps, depending on some conditions, this is managed by a flow-element that we call a *Control* (GSDEM 2015). Figure 3 shows the flow mechanism of a process step appearing in planning, in developing the IT system and in production with phase dependent elements of the process step. The terms *"method module"* and *"supporting module"*, familiar from Figure 2 are used also here, but now abbreviated.

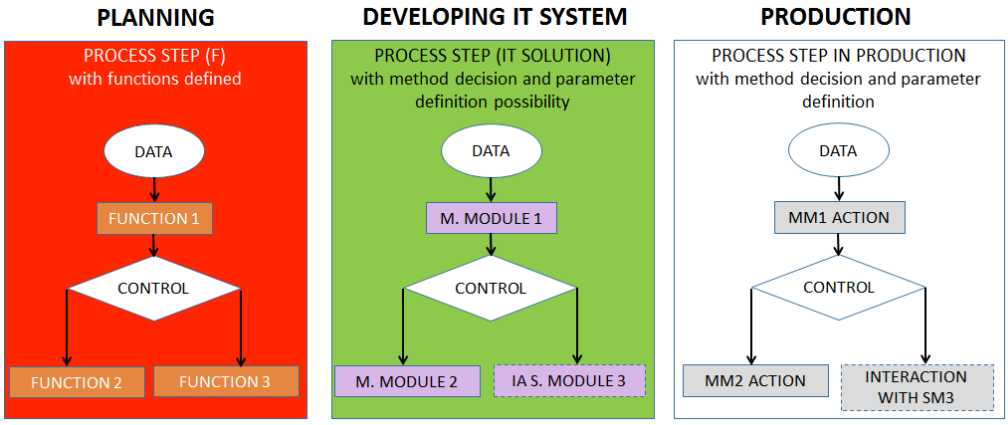

Figure 3. Process step in planning, developing IT system and production

24. These process step flow charts are concentrated on functions, method modules and actions with modules. However, one crucial part of production is not visible in the charts. The process flow and the process steps cannot be conducted in practice without some *process metadata*. The methods chosen for functions may produce data (*output metadata*) which is necessary for subsequent phases in a process step or the process flow (necessary as *input metadata*). In addition, the parameterization decided (in whatever form that is) is also *input metadata which steers the process* at different levels. Also other types of metadata not strictly affecting the production exist in the editing process. See GSDEM (2015) for more detailed description of the topic.

25. The functions in all three function types (*review, selection* and *amendment*) can be expressed in such levels which produce *indicators* "in terms of their effect on data, performance, expense, and so on, which are independent of the chosen statistical method or implementation thereof" (Pannekoek et al. 2013). These indicators can be edit violations, scores, selections for amendment etc. In most cases the indicators are produced by modules (or other solutions) when a specified method for a function is carried out.

26. An exception in creating the indicators is the possibility to *calculate amendment indicators by using data sets before and after the amendment*. For interactive operations the supporting module can register the indicator value, though this practice is rather rare (e.g. manual actions in the application might not leave "traces" into the system except perhaps the amended value itself).

27. The indicators might not be in a form of variables in some cases. For example, when a classical "IF+THEN" rule is used, one recognizes the error with an edit rule, decides to carry out a specific correction to the observation with edit rule violation and finally makes the correction by a defined function. No indicator variables are made in this case, and the only outcome is the altered variable value.

28. Although the parameters and methods might be spread in various program codes, they still are metadata, though not in a very organized form. In practice some of those process steps or methods with parameterization may be conducted with the editor updating, revising and "pressing the button" for programs.

29. At best the IT system can include specific parameter data sets with possibilities to apply them in different tasks (in modules) and to utilize them for evaluation and documentation. In some cases also the set of methods for different purposes is available as metadata.

III. Monitoring and evaluation of editing leading to improvements in editing process

A. Statistical data, editing process and results: targets of monitoring and evaluation

30. How do we know when we have a good editing process going on in order to study and correct the statistical data and to obtain reliable results? It may require the evaluation of all these three elements during editing process: *statistical data* and *results produced in various phases of the process* and the *process itself*.

31. When the editing process is going on, we may like to *monitor the editing process* in order to get information about different aspects of the process and the statistical data and results in that process. The *evaluation of the editing process* is a larger concept and it contains different types of evaluation actions, not necessarily conducted during the process.

32. We know about the editing process and its functionality in statistical data for the most part *by carrying out the process using statistical data*. This can happen either *in production* or with *some test data set(s)*, usually data sets from earlier rounds.

33. The process can be monitored or evaluated either in a *measurable* or a *non-measurable way*. The measures can be directed to the statistical data or the process and its progress. The process part requires a metadata system to provide necessary information about the process.

34. In some contexts these measures are called *quality indicators* (e.g. EDIMBUS 2007 and Ollila et al. 2012), although many of the measures are descriptive and "quality" when referring to accuracy and effectiveness is not necessarily measured. Furthermore, the indicators produced in functions are sometimes also referred as quality indicators e.g. in Pannekoek et al. (2013). When describing the process, alternate expressions could be e.g. *process indicators* (Lindgren and Odenkrantz, 2014) or *process metrics* (GSDEM 2015).

35. The latter, non-measurable class of evaluation is a *heterogeneous set of different observational or analytical efforts* to evaluate the editing process and/or its statistical data and results. The usual way is to evaluate by experience in production or test situations.

36. A specific type of evaluation appearing occasionally is to *study a suspicious phenomenon concerning the statistical data* observed during the editing process. The aim is usually to reveal new sources and reasons of systematic errors in the statistical data. This may include also some analysis of the data.

B. Improving editing process

37. When we want to improve the editing process, there are several parts in the process to which the improvement can be targeted. The next categorization illustrates those parts and it is based on process levels appearing in GSDEM (2015) and the IT system realizations of these levels from bottom to top.

38. **Changes in parameterization**. Generally at least some changes in parameterization are carried out between different rounds, some of them following the changes in the substance area of the statistics. Probably the most common example is to adjust edit rules or parameters of rules (limits, conditions etc.) or to make new rules. Here the exact defined edit rule is considered to be a parameter, but in some cases e.g. the type of edit rule could be interpreted as a method (e.g. range of valid values, comparison to a historic value).

39. **Changes in method selections**. The methods are changed occasionally in the process steps, mainly due to some new reasoning or studies in tests (e.g. changing imputation from mean imputation to median imputation or applying more complex cluster analysis in error recognition) or when there are new IT modules available for more sophisticated methods.

40. **Changes in method modules**. The modules (procedures, macros, program codes etc.) for methods are subject to development especially in statistics having non-systematic structure of the IT system. Sometimes the needs for new methods require new module solutions or even tool packages.

41. **Changes in functions**. There may be changes in functions appearing in the process steps, when some renewal is carried out in parts of the statistical process. An example of this is the inclusion of the selective editing process step and its functions to the statistics in the development project of Statistics Finland.

42. **Changes in process steps**. The process steps can change (or new steps can emerge) if there is a need for revision in some part of the process flow, as in the example above.

43. **Changes in systems carrying out process steps**. The substantial changes in the IT system are quite rare, but they are conducted especially when there is a need for more systematic processing system with all-covering metadata structure and calculation of indicators at different levels. An example of this is the development of the EG EDIT application for reproducing the phases of the editing model, applied in some statistics at Statistics Finland (Oinonen 2014, Ollila et al. 2012).

44. **Changes in process flow**. The changes in the process flow of editing in the statistics are rather exceptional, and they are usually tied to large-scale projects in order to improve the efficiency of the editing process, possibly following the idea of harmonizing the editing process among the statistics in the statistical office.

IV. Evaluation and improvement of editing at Statistics Finland

A. Background

45. During recent years three projects dealing with editing were conducted at Statistics Finland. The projects have produced applicable solutions in the perspective of editing process as well as good practices in editing. The SAS EG application called EG EDIT was developed for the needs appearing during these projects. The project proceeding now has a target to implement selective editing in some chosen statistics, and this work has provided a lot of interesting results in the field of editing. Oinonen (2015) describes the situation and also provides results and experiences in more detail in her paper.

46. In this chapter the results of the projects and the development of EG EDIT are studied in the context of evaluation and improvement, utilizing the definitions and classifications presented in Chapters II and III. Also challenges and difficulties appearing in the development process are dealt with to some extent. The subtitles of this chapter follow the seven-class categorization provided in Section III B, but this time the order is from top to bottom.

B. Changes in process flow

47. Normally the statistics won't be changing the process flow of their editing unless there is a specific project for that purpose or some renovation in the overall process structure of statistical production.

48. The project opened the possibility to revise the process flow structure, especially including the selective editing process step in it. Although the process flow in statistics might not have been clearly defined in some cases, the supportive review sessions with reports on main practices gave sufficient information to the team for further development of the process. Including some new or altered steps to the process required some adjustment steps for the surrounding, possibly different IT environment (see Oinonen 2015).

C. Changes in systems carrying out process steps

49. A SAS EG application called EG EDIT was constructed to utilize the properties of BANFF (Banff Support Team 2007) and SELEKT (Norberg et al. 2010) packages with additional macro modules for the needs in designing and implementing the editing process of various statistics. The package has a metadata system collecting necessary information about the process and definitions for the process. EG EDIT has been demonstrated in detail in Oinonen (2014). As an alternative, EG EDIT has an important role in changing systems carrying out process steps.

50. Concerning process metadata, the indicators produced by methods chosen for functions of different type (review, selection and amendment parts) are available in the status data sets (familiar from BANFF) in different phases of the process (in test version also amendment included). The definitional metadata for steering the process (collected automatically from the definitions given in EG EDIT) is available in full during the process. The parameterization of the current process is one part of the definitional metadata.

51. One feature of the process-like project form in EG EDIT is to create process indicators automatically when the implementation of the application is going on. At this point these indicators cover aspects of item response rates, edit rules in various forms of study (from BANFF) with additional proportion tables, descriptions about the preliminary automatic corrections and their effects, overall and categorized occurrences of error alarms via different methods (edit rules, selective editing, outlier

recognition). When the current edited data is available, the occurrences of successful edit rules (finding error) by variable can be studied and measures describing differences between unedited and edited data sets. More indicators are supposed to be constructed in EG EDIT, if resources and time could be allocated for that purpose.

D. Changes in process steps

52. A process step added to the process flow in some statistics was the *automatic correction of observed fatal errors with exact solutions*, mainly dealing with thousand errors. In order to create methods it required additional studies with unedited and edited data sets based on some knowledge on substance in advance. The tests were carried out in EG EDIT, which has tools for studying edit rules and their effect to the estimates.

53. A more complex process step with new functional elements for many statistics was *selective editing*. This phase was the key reason of the latest project and several studies with various ways were conducted in order to reach sufficient solutions for each statistic under study. See Oinonen (2015) for more description.

E. Changes in functions

54. The new process steps presented above required function definitions in those steps as well, e.g. score calculation and error list production, though the latter appeared in some statistics defined in a different kind of process step. Usually the flow in that process step was considered to be same from one case to another.

F. Changes in method modules

55. EG EDIT has a set of modules in a constant basic structure and the modules are steered with separate definition blocks of parameters (Oinonen 2014). At the moment there are no plans for changes in that module pattern, albeit improvements in functionalities and new modules could be expected later in the future.

56. An exception in this situation is the *continuing development work on error lists*. The error list provided by SELEKT (Norberg et al. 2010) is emphasizing methodological measures of the scoring process. The statistics have varying desires about what kind of information the error lists to be applied in manual studies should contain. For these *descriptive error lists* some new module solutions are created.

57. Some modules (or programs) are added into the EG project in order to prepare data sets, possibly from another environment, for the process in EG EDIT, and on the other hand, some data is prepared for purposes in another environment. These features are important especially when *doing simulation of the process during data acquisition* or *preparing the system for production* (see Oinonen 2015 for illustration).

58. It is not self-evident that the modules will work in all data situations and definitions. For example, in two-stage observation situations the lowest level elements can be at a too detailed level for proper treatment in the system and utilization in studies, e.g. as in *Waste statistics* with waste codes see (see Oinonen 2015). Some adjustments (e.g. aggregate calculations to a more robust level) might be made in advance in order to prepare a challenging data situation suitable for further processing. This emphasizes the need for technical improvements in challenging data cases.

G. Changes in method selections

59. Before different editing projects in recent years the set of selected methods for the editing process was rather small in statistics at Statistics Finland, with some exceptions. Though still not having a vast variety of methods, new knowledge, available tools and possibilities to carry out more complex methods rather easily have enabled both method changes and new method selections for new functions (tasks) in the process.

60. During the latest project the testing processes for selective editing carried out in EG EDIT with the properties of SELEKT were essential to find effective edit rules and test variables with functions with suitable methodological parameters (see Norberg et al. 2011). The role of review discussions at the beginning together with a form asking key questions on editing in that statistics was essential to get into the practices which should be taken into account when developing selective editing (Oinonen 2015).

61. For selective editing some measures for evaluation of the process were used. They were based on some pairs of unedited and edited data sets, producing *pseudo-bias* tables for finding efficient edit rules and edit functions (via test variables, see Norberg et al. 2011). These SELEKT-based tables were also supported by an additional comparison with the partially edited and edited overall estimates at different percentage levels of manual editing.

62. *Finding suitable types of edit rules* was an important part in developing efficient error recognition in statistics. In addition to rules found based on error studies on suspicious phenomena in the statistical data, "tricks" for efficient error recognition in some cases were found. For example, in surveys when we know that in some places in the questionnaire some respondents can have a "tradition" to fill the answers wrongly, using the change indicator from the previous unedited and edited data may find surprisingly many errors in the unedited data. The data structure in EG EDIT supports this kind of study.

63. To support these studies, EG EDIT has now an additional module for studying edit rule hits and successful recognitions of errors in them. The same module also shows which errors are not found by the existing edit rules in order to find new rules for those.

H. Changes in parameterization

64. In normal production the parameter changes are rather modest, and they might be based on evaluation of process indicators describing the edit rules and other methods, available in EG EDIT when executed. When e.g. using simple limits in query edits or parameters for outlier recognitions, some monitoring of the development in the field to be studied is considered to be good practice, and there are plans to build this kind of mechanism in the process of some statistics.

65. The parameterization while developing the editing process of the statistics is rather frequent, especially exact definition of the edit rules. This work needs substance knowledge of the statistic in question.

66. EG EDIT has parts for parameter definition without a need for programming. However, the transfer of the edit rules from the old system can be surprisingly laborious, and it may contain problems in interpretations or adjustment problems. The locations of programs containing edit rules can sometimes be difficult to be found. The data sets available may not have all variables needed in the edit rules.

I. Development of editing and EG EDIT in future

67. The process of developing editing is ongoing, and some statistics are already interested in making their system to follow the structure of EG EDIT with selective editing. However, at the moment it seems that separate projects with sprint periods won't be conducted in the near future, and the development might be done only among normal routines of the methodological unit, adjusted to diminishing resources.

68. On the other hand, there might not be extra time available for developing EG EDIT in a few years' time. A lot could be done to the flexible SAS EG system in order to make the application consisting a wider range of methods and properties (especially regarding indicators in time), but there won't be much spare time for that purpose.

69. Although bugs have been corrected, EG EDIT needs revision a lot in order to make it more usable (especially when creating edit rules), to get rid of quick solutions for some challenging situations and to make the system more responding to various kind of error situations and mistakes in definitions,

which can easily be rather frequent. At the moment we'll manage well with EG EDIT, but some prior experience on difficulties appearing in the project helps a lot to avoid to be stuck in error situations.

References

Banff Support Team (2007). Functional Description of the Banff System for Edit and Imputation, Statistics Canada.

EDIMBUS; Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B. &, Kilchmann, D. (2007): *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*, EDIMBUS project report.

GSDEM: Task Team on a Generic Process Framework for Statistical Data Editing - Generic Statistical Data Editing Models (GSDEM), draft prepared for the 2015 UNECE Work Session on Statistical Data Editing.

Lindgren, K. and Odenkrantz, M. (2014). Applying process indicators to monitor the editing process, Work Session on Statistical Data Editing, Paris, France, April 2014.

Norberg A., Arvidson, G., Kraftling, A. & Nordberg, L. (2010). SELEKT - A Generic SAS™ System for Selective Data Editing, Statistics Sweden.

Oinonen, S. (2015). Implementation of selective editing methods at Statistics Finland using innovative and efficient team work methods, Work Session on Statistical Data Editing, Budapest, Hungary, September 2015.

Oinonen, S. (2014). SAS Enterprise Guide project for editing and imputation, Work Session on Statistical Data Editing, Paris, France, April 2014.

Ollila, P., Ahti-Miettinen, O. & Oinonen, S. (2012). Outlining a Process Model for Editing With Quality Indicators, Work Session on Statistical Data Editing, Oslo, Norway, September 2012.

Pannekoek, J., Scholtus, S. and Van der Loo, M. (2013). Automated and manual data editing; a view on process design and methodology, Journal of Official Statistics.

Pannekoek, J. and Zhang, L.-C. (2012). On the general flow of editing, UNECE Work Session on Statistical Data Editing, Oslo, Norway.

Saint-Pierre, E. and Bricault, M. (2011). *The Common Editing Strategy and the Data Processing of Business Statistics Surveys*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Vale, S. (2011). *The Generic Statistical Business Process Model and Statistical Data Editing*. Invited paper in UNECE meeting, Ljubljana, Slovenia.