WP. 14
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (ii): Global solutions to editing

The development of a data editing and imputation tool set

Claude Poirier, Statistics Canada

I. INTRODUCTION

1 The development of global solutions, whether national or international, should start small, with realistic objectives. In this context, cost efficiencies would focus on global needs rather than local ones. For data editing and imputation, this translates into the robustness of sound and well tuned approaches that would perform well with a wide variety of data sources. The paper identifies common requirements for both statistically collected data and administrative data. It puts into perspective dimensions of quality in order to evaluate potential solutions for the global objectives. It suggests methods that can be considered as default choices in practice, providing the starting points for assembling a tool set. It also defines and addresses the interoperability issues in a multi-platform context.

2 The proposed tool set is far from being the result of a complete and exhaustive evaluation by an independent authority. More investigation would be required on several software packages to complete this evaluation. For instance SEE – the Significance Editing Engine developed by the Australian Bureau of Statistics – would be worth investigating. Similarly, lessons may be learned from the New Zealand platforms POSS for social statistics and BESt for economic statistics.

3 Section II lists the international initiatives in order to prepare the reader for Section III related to the desired features of an international tool set. Then a very basic tool set is proposed in Section IV followed with a description of functional gaps yet to be filled in Section V.

II. INTERNATIONAL INITIATIVES

A. HLG-BAS vision and actions

4 The High-level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS) was created by the Bureau of the Conference of European Statisticians in 2010 to reflect on and guide strategic developments in the ways in which official statistics are produced. The group documented its vision the year after (see HLG-BAS 2011a). In addressing issues related to statistical products, it was noted that statistical organizations have been conditioned to look at data through a filter that discards all data without the stamp "officially sourced" or "officially collected" as generally not fit for our use. There was little of that type of data anyhow. This traditional approach is now rapidly becoming outdated. Indeed, Internet companies and other entities have demonstrated that something very interesting is happening. We are in a changeover from a society with little or no data available to one that has an abundance of data. In this light we have to rethink our traditional business values.

5 With respect to processes, the vision states that changes in our society increase the need for more and quicker statistics. Quality becomes negotiable but needs to be communicated. The challenge for statistical organizations is to be sufficiently flexible and agile so as to provide statistics according to user

needs, at an acceptable cost. Statistical organizations have considerable experience in harmonising statistical products and regulating requirements within the different statistical domains. Providing statistical results has been regarded as statistical production for decades, but international cooperation has not concentrated much effort on harmonising production processes, leaving many possibilities for increased standardisation.

6 The production of statistics should be based on common and standardised processes. In some cases, standardisation of processes and the availability of international data sources could lead to statistical production at a multi-national level. The HLB-BAS proposed some actions during its October workshop (HLG-BAS, 2011b). The strategy document states that the statistical organization should streamline processes within a more industrialised environment. This paper addresses this challenge in Section IV.

B. Statistical Network

7 The Statistical Network was created around the same time as the HLG-BAS. It is an initiative from the Australian Bureau of Statistics in order to set the road to industrialisation and standardisation. In June 2010, members of the Statistical Network met in Paris to identify key opportunities, and the industrialisation of editing was one of them (Statistical Network, 2010). The group responsible for the industrialisation of editing started to think about a plan and some outcomes. The proposal was to use existing resources such as the Recommended Practices Manual (ISTAT, CBS and SFSO, 2007) that was developed in the framework of the European Project "Editing and Imputation in Cross-Sectional Business Surveys (EDIMBUS)", as well as some editing and imputation standards and guidelines that were developed by National Statistical Institutes (NSIs). The recommended method will then drive the identification of a minimum set of standard methods and IT tools for editing and imputation. Unfortunately, the industrialisation of editing has been put on hold for now within the Statistical Network. Nevertheless, this paper proposes some tools based on EDIMBUS principles.

8 On top of recommended methods, the editing and imputation principles that have been defined within the scope of the industrialisation of editing (Zhang et al., 2011) must be considered when proposing a tool set. These principles include the following 7 aspects.

i. The original data provided by the respondent or data supplier should be maintained, wherever possible;
ii. All outputs must have a documented editing and imputation plan;
iii. Error prevention must be performed to ensure that data are processed intelligently given the concepts that are being measured;
iv. Editing and imputation methods must be chosen carefully, ensuring they align with international best practices;
v. Focus must be put on important anomalies;
vi. The editing process should be automated where possible, with the aim of minimising manual intervention;
vii. Information about quality must be captured, analysed and shared amongst users and stake-holders;
viii. All staff working with data collection must have access to documentation of, and training in, editing and imputation principles, methods and procedures.

III. DESIRED FEATURES OF A TOOL SET

A. Desired Functionality

9 The EDIMBUS Recommended Practices Manual (ISTAT, CBS and SFSO, 2007) listed a series of recommendations for business surveys. Although these don't prescribe specific methods, it implicitly suggests some methods while recommending related principles. These methods most often apply to numeric variables as typically found in business surveys. They are related to either the editing process, the imputation process, or the estimation process. Poirier (2011) listed the Canadian requirements as they relate to editing and imputation. These requirements include the EDIMBUS implicit methods, in addition

to a few more that are often related to categorical variables as typically found in social surveys. The desired functionality in the proposed tool set is presented in Table 1 (section IV), and reflects the EDIMBUS recommendations.

10 Editing process: EDIMBUS recommendations on editing are related to on-line edits, fatal edits, distribution edits (most often not fatal), influential edits, outlying edits, selective editing using scores, deductive or deterministic edits, Fellegi-Holt minimum change principle, and macro editing most often during certification activities.

Canadian requirements consist of the above methods in addition to flow edits, conditional edits, decision logic tables, validation of edit rules, editing cells or groups, and edit failure status.

11 Imputation process: EDIMBUS recommendations on imputations are related to imputation cells or groups, rule-based imputation, deductive imputation, model-based imputation, and donor-based imputation.

Canadian requirements consist of the above methods in addition to imputation by prorating.

12 Estimation process: EDIMBUS recommendations on estimation are simply related to the computation of the variance due to imputation. This is because the manual focuses on aspects of a typical survey cycle as long as they have a relation with the editing or imputation processes.

Canadian requirements consist of the above methods, with nothing else.

B. Efficiency

13 Organizational efficiency is already one of the priorities of several NSIs. It becomes mandatory to look at what is available off-the-shelf before initiating a costly development project. Furthermore, efficiency becomes the driver for the vith principle listed above (Section II.B). As of now, increasing efficiency requires that NSIs be innovative in their business practices. Processes must be examined from various angles: how information is managed, how systems are built, how programs are governed. In this context, all assumptions are questioned and greater emphasis must be put on global (corporate departmental-level) efficiency, as opposed to the historical local (survey-level) efficiency. The result of standardization and centralization of processes is that specific surveys may lose some flexibility for the best interest of the corporation as a whole. An international initiative that aims for standardized methods and tools can only help the individual NSIs meet their own objectives with respect to efficiency. NSIs should simply establish their individual risk management strategy to make sure that their strategic outcomes are met.

14 For several decades, many NSIs had let survey managers initiate system development without exerting too much control. When agencies had to face budget cuts, they started to streamline technologies and functions. Many major system redesigns were then initiated across NSIs. The first step of such projects is the identification of goals and objectives. Fortunately, NSIs don't develop statistical tools with the intention of selling them. They rather target the automation of their current processes by borrowing or acquiring existing products whenever they exist, and by developing some others only if needed. In the latter case, the developers require a clear understanding of the user requirements. In an international context, the idea consists of combining requirements of several NSIs and taking advantage of common efforts. This means that challenges can be assigned amongst several developers and development time can be reduced resulting in a more timely and robust product for the investment.

15 A first step in this quest for efficiency is to identify the tools to be targeted by most NSIs' programs. It is important to stay modest and realistic at the beginning, this by harvesting the low hanging fruit. Therefore, we should focus only on methods and tools that are currently used by NSIs, and ideally the ones that are already recognized by the statistical community as sound and well tested. Some quality criteria must be set in order to justify the choice of specific tools against some others.

C. Quality Criteria

16 Several NSIs put in place quality manuals which guide managers through quality issues associated to survey steps. Statistics Canada Quality Guidelines (STC, 2009) is one of these. The document mentions that "A significant feature of the management of quality is the balancing of quality objectives against the constraints of financial and human resources". The goal of a software development project should not be the maximization of quality at all costs, but the achievement of an appropriate balance between the quantity (the number of features) and quality.

17 A multidisciplinary project team is required to conduct a system development project. The project team makes the many decisions necessary to ensure that there is an appropriate balance between concerns for quality and considerations of cost. The team should consider all of the dimensions of quality, even if some are conflicting, while identifying the optimal system functionality. The following paragraphs describe the various dimensions of quality according to STC (2009). While these dimensions primarily apply to statistical products, they can be adapted for system development.

18 Relevance: The relevance of a processing system reflects the degree to which it meets the real needs of the client in terms of functionality. In the case of multiple clients, relevance is influenced by their varying needs. Within the context of statistics programs, the relevance of systems is the key dimension of the quality. A system may be reliable, robust, powerful, user-friendly and well documented, but if it does not address the needs of the statistical program, it loses its quality. The challenge of an agency is to weigh and balance the needs of current and/or potential users with the resource constraints. This exercise requires the translation of user needs into program approval and budgetary decisions. To achieve relevance, an agency should care about system diversity. In other words, it should try to keep the duplication of functions as low as possible across systems, and prioritize development initiatives which address system gaps.

19 Accessibility: The accessibility of a system refers to the ease with which users can work with it. This includes items like platform constraints or portability, installation protocol, flexibility, adaptability, integration, power, user-friendliness, documentation, support resources, etc.

20 Interpretability: The interpretability of a system reflects the level of information that is made available to help users in understanding both the system input and output. Systems require some sort of data dictionary or metadata to ensure interpretability, regardless of the complexity. Other elements which contribute to interpretability include system and methodology documentation, record layout, on-line help, diagnostic reports, error and warning messages, etc.

21 Coherence: The coherence of systems refers to their level of standardization and interoperability. It is the key element of integration. It helps to bring together systems to form a suite of systems which can be run one after the next to ideally process all survey steps. Two points of view must be considered on this matter: the developer and the user. For the developer, system coherence is dependent on the existence of a framework. This includes the consistent use of operating systems, foundation software, modules, function calls, programming objects, etc. With respect to the user's point of view, system coherence is achieved through the consistent use of graphical objects, concepts, variable names, data formats, code sets, etc. In terms of data structure, this means that the output of one system becomes the input for the next system. The coherence across systems directly helps users to shorten their learning curve, and it usually reduces the maintenance and support cost for the whole suite of systems.

22 Accuracy: The accuracy of a system is the degree to which its outcome corresponds to what it was designed to process. Accuracy is ensured by a quality assurance process during the development phase that includes several phases of testing.

23 Timeliness: Timeliness can be seen from different angles. During the development phase, it refers to the delay between the date when the system is required and the date when it becomes available. During the practical use of the system, timeliness mainly refers to how fast it runs and whether or not it meets the user requirements in terms of performance. Like statistical products, the timeliness of systems influences their relevance.

C. Important Software Characteristics

24 In pooling requirements together, the agency should focus on the use of a single data structure for most of its systems. For example, some NSIs steered their system development toward SAS. There are several reasons for this, including the fact that there is already widespread usage of SAS in the statistical community and thus a broad knowledge base already exists. SAS also satisfies the user requirements for portability across the different computing platforms.

25 As opposed to commercial developers, the mandate of statistical agencies is not to develop elegant systems but to support statistical programs. They must satisfy the internal needs by providing a simple and robust infrastructure. This means that the focus should be on the processing aspect in order to make the systems reliable, efficient, affordable, and robust, require minimum maintenance, cheap and coherent with the infrastructure. While the ease of use and reuse clearly competes against the above features, from the user's point of view it is important that the learning curve be short in order to speed up development of survey applications. User-friendliness of systems is one of the objectives to be strived for in reaching the overall goal.

26 As mentioned by Deguire, Reedman and Wenzowski (2011), the application of sound software engineering practices must be supplemented with a clear understanding of the outcome of the process. In the context of generalized systems and global solutions, it can be described with four characteristics that the resulting software must possess.

27 Adaptability is the ability to adapt to different requirements. A global solution, as its name implies, is to be used by a large number of surveys and, as such, must be developed to accommodate runtime processing specifications. To achieve this, systems should be developed in a highly cohesive modular fashion whereas each module implements a specific statistical function. By not overloading a module, it is easier to build the software in a flexible manner such that the user can easily alter the constraints and the assumptions. In other words, the module must be 'parameter driven' or 'metadata driven'.

28 Reliability is the ability to produce accurate results in a timely fashion. The accuracy and timeliness aspects were already addressed in Section III.B above. Because of their status as international standards and their foreseen widespread use, global solutions must implement sound, well understood and defensible statistical methods. As such, a global solution is not a playground for research. Users expect the software to be robust and that its execution produces results that can be trusted. With respect to timeliness – or efficiency - it should not be an afterthought. It starts at the specification stage by questioning inefficient methods. Statistical researchers play a key role by prototyping those methods before a decision is made on their implementation.

29 Maintainability is the ability to enhance existing functions or add new functions, and to adapt to new operating environments. The development of generalized systems is a long and expensive process. The resulting software must be in production for many years to justify such an important investment on the part of a statistical agency. To do so, the software must be built in such a way that it will survive many changes in the operating environment and be able to entertain enhancements to its functionality so that new relevant methods can be easily implemented. Again, the software should be built as a collection of highly cohesive modules. Each of these modules can be enhanced as long as the enhancements pertain to the statistical function it implements. New statistical functions should be built as new modules and hence augment the collection. The software should also be built with a layered approach in order to isolate core components from the operating environment. The adoption of an application virtual machine such as SAS® or Microsoft .Net is critical in this regard.

30 Interoperability, as referred to by the coherence aspect in Section III.B above, is the ability to interoperate with other systems and software. Several generalized systems modules are typically assembled to implement a specific application for a survey. Furthermore, custom and commercial software typically supplement the generalized system modules. This reality implies that these must not only be cohesive but also loosely coupled so that they don't have interdependencies with other modules. The modules must have a clean and well-defined programming interface. This interface allows the output of a module be the input of another with no knowledge of their respective internal implementation.

IV. THE BASIC TOOL SET

31 This section proposes a few tools and justifies their inclusion into the tool set given the criteria and desired features listed in Sections III.

A. Banff

32 Banff was developed at Statistics Canada and became the recommended editing and imputation system for Canadian business surveys. It was initially interacting with Oracle tables, making it heavy to run, before being migrated into a more flexible and user-friendly SAS environment. It performs automatic imputation while respecting specified edits. The system applies linear programming techniques to conduct the localization of a minimum set of fields to be imputed, and then various algorithms are used to perform the imputation. The processing is entirely driven by linear rules defined on numeric variables. Univariate outlier detection is based on quartile distribution. Outlying observations can be flagged for imputation or simply excluded from subsequent calculations. The imputation function offers three imputation methods: Deductive, Donor, and Estimators. Deterministic imputation identifies cases in which there is only one possible solution given the edit rules. Donor imputation uses data from the nearest neighbour and the estimator function provides a wide set of models using historical or current information.

33 **Justification**: In terms of efficiency, the vast usage of Banff across several NSIs proves that it can be trusted. It can satisfy most business survey applications and its modular approach requires only a modest investment from the part of the users to integrate it into their survey environment.

34 With respect to the quality criteria, Banff meets all the six dimensions listed in III.B. Its sound methods make it very relevant for editing and imputation of business surveys and administrative financial data. Its SAS procedures ensure its accessibility for most NSIs. Its documentation and diagnostic reports meet the interpretability criteria. Its integrated functions as well as its flexible SAS environment enable the coherence with other processors in the survey chain. Its several methods and options as well as its long history in production environments demonstrate the accuracy of its functionality. Finally its current state and its performance meet the timeliness criteria.

35 As for the four important characteristics, the adaptability of Banff is ensured by its highly cohesive procedures that can be called in various orders to satisfy survey requirements. Its reliability is shown by the international experience for more than a decade and by the level of support that is available for the system. The maintainability is ensured by its intrinsic structure and the support team which stays available to modify and expand functions. The interoperability is demonstrated above through the coherence criteria and its flexible SAS environment.

36 Banff functionality is presented in perspective with other proposed components of the basic tool set, in Table 1 below.

B. CANCEIS

37 The Canadian Census Edit and Imputation System (CANCEIS) was also developed at Statistics Canada. It was initiated to handle the categorical variables of the 1996 Census of population. Its functions were expanded to process a mixture of categorical and numerical census variables. This now makes the system suitable for some social surveys. It can perform both deterministic and donor imputation. Its goal is to minimize the number of changes while making sure the imputation actions are plausible. It uses edit rules defined by decision logic tables to identify records that need imputation and records that can be used as donors. For each record to be imputed, the system tries to find the nearest record that can be used as a donor and identifies the optimal imputation action by borrowing as few donor fields as possible to impute the record in error. CANCEIS is intrinsically different from Banff since it finds the donor before identifying the minimum imputation while Banff does the opposite. The two systems deserve a place in a generalized tool set because one would not perform well in a purely numerical context and the other cannot process categorical variables.

38 **Justification**: In terms of efficiency, the vast usage of CANCEIS across several NSIs proves that it can be trusted. It can satisfy most census applications. The development of an application may require non negligible investments, but keeping the size of a census project in perspective makes the system very affordable for its users.

39 With respect to the quality criteria, CANCEIS meets all the six dimensions listed in III.B. Its sound methods make it very relevant for editing and imputation of census categorical data. Its foundation software C ensures that it can be compiled on any IT platform, ensuring the accessibility for most NSIs. Its documentation and diagnostic reports meet the interpretability criteria. In terms of coherence, CANCEIS offers the desired standardization but doesn't directly offer the connectivity that one would desire. On the other hand, simple pre-processors are easy to build to ensure its program specific connectivity. Its several options and its long history in production environments demonstrate the accuracy of its functionality. Finally its current state and its performance meet the timeliness criteria.

40 As for the four important characteristics, the adaptability of CANCEIS is ensured by its highly cohesive modules that showed its adequacy for several surveys and censuses around the world. Similarly, its reliability is shown by the long international experience even if international support is not yet available. Maintainability is ensured by its intrinsic structure, its documentation and its support team which stays available to modify and expand functions. The interoperability is the only characteristic that is not demonstrated for CANCEIS.

41 CANCEIS functionality is presented in perspective with other proposed components of the basic tool set, in Table 1 below.

C. Selekt

42 As mentioned by Statistics Sweden (2011), selective editing emphasizes the search for significant errors, accepting that resulting data sets would contain a number of errors but with negligible impact on the statistical product. In this context, they developed Selekt – a selective editing system – with the objective of reducing the collection cost. They take the right approach in looking for the source of error to improve the survey process rather than to correct all errors. One of their principles is to prioritize known respondents because their historical values can drive the selective editing. Selekt measures the suspicion level, the potential impact of the related correction, local scores for each suspicious data point, global scores (as the max of local scores) for each record, and a pseudo-bias measure in order to drive the process. This is done with some control on the importance of variables, the level of aggregations, and the resulting cells within tables (or estimation domains). Since the real impact is always unkown, Selekt generates replacement values for each suspicious data point, using either longitudinal (micro level) techniques or auxiliary and current data. For efficiency and accuracy reasons, the process is run on groups of records, given their similarities.

43 **Justification**: Although it is relatively young, Selekt is already being used across several NSIs, and some others are planning on using it in the near future. Statistics Canada is one of them. Selekt can satisfy most business survey applications which deal with numerical data. Its growing community already makes it an efficient choice for selective editing.

44 With respect to the quality criteria, Selekt meets all the six dimensions listed in III.B. Several evaluation studies have proven its sound and relevant methods for business surveys and administrative financial data. Its series of SAS macros and SAS environment ensure its accessibility for most NSIs. Its system and user documentation meets the interpretability criteria. Its SAS data structure offers enough flexibility to interact with other SAS systems, such as Banff, this to satisfy the coherence criteria. Its several features as well as the various independent tests demonstrate the accuracy of its functionality. Finally its current state makes it usable in production environments – a timely solution.

45 As for the four important characteristics, the adaptability of Selekt is ensured by its highly cohesive macros that can be called with inter-processors in between, or can be replaced with some other modules. Its reliability is shown by its quick adoption by several of the most advanced NSIs. The

maintainability is ensured by its intrinsic structure. Although no international support is offered, Selekt's team considers modifications and expansions to satisfy users' needs. The interoperability is demonstrated above through the coherence criteria and its flexible SAS environment.

46 Selekt functionality is presented in perspective with other proposed components of the basic tool set, in Table 1 below.

Table 1. Functionality of the proposed tool set

EDITING	Banff	CANCEIS	Selekt
On-line edits			
Construction of groups			✓
Editing within groups	✓	✓	✓
Fatal edits	✓	✓	✓
Distribution edits	✓		
Outlying edits	✓		✓
Selective editing (with scores)			✓
Deterministic edits	✓	✓	
Fellegi-Holt (minimum change)	✓	✓	
Editing of macro data			
Graphical editing			
IMPUTATION	**Banff**	**CANCEIS**	**Selekt**
Imputation within groups	✓	✓	
Rule-base imputation		✓	
Deterministic imputation	✓		
Model-based imputation	✓		
Donor-based imputation	✓	✓	
Prorating imputation	✓		
ESTIMATION	**Banff**	**CANCEIS**	**Selekt**
Variance due to imputation			

V. FUNCTIONAL GAPS

47 From Table 1, it becomes clear that a few requirements are not satisfied by the proposed tool set, editing of macro data (or aggregated data) and graphical editing being two of them. These functions would provide prioritization and guidance for micro-editing. Furthermore, they would be useful during the certification phase of any statistical products. A few systems were already produced by NSIs. An evaluation according to the several criteria listed in this document is suggested.

48 Another limitation is the lack of on-line editing functions. In a typical statistical process, the edit functions are used to identify suspicious records with suspicious data values, and bad records with data values that need to be fixed. Then a series of investigation and manual interventions are prioritized through selective editing in order to fix the most important errors before automated corrections are performed on the yet unresolved errors. Ideally, the tool set would handle these typical steps while sharing common editing routines amongst steps. This means some routines should be available from a batch environment (for instance for the initial batch edits) as well as from an on-line environment (for instance to be run on individual records during the collection process or the manual editing process). This is unfortunately not possible with the current tool set.

49 A third item on the wish list is the variance due to imputation, a key function for typical business surveys, especially when selective editing is applied together with a mass imputation strategy for less

important units. Statistics Canada is working on a new estimation package, G-Est, which will offer variance due to non-response and imputation by January 2014. It is based on the methodology of Särndal (1992) and its extensions implemented in the prototype system SEVANI (Beaumont and Bissonnette, 2011). The development work has been initiated to make Banff's output compliant with the input needs of the variance due to imputation.

50 Finally, a standard micro format would be very useful when the tool set is in place, this lot let processes to interact with each others. Zhang et al. (2011) said that "It will be very helpful for developing tools of database management, if there is a standard concept of data on the micro level, which we refer to as the common micro data format". Database management tools would help the connectivity feature, even if SAS is contributing to this feature for now. The Generic Statistical Information Model (UNECE, 2012), once completed, will guide this aspect of development.

51 Note that several other editing and imputation functions are also missing from the functionality of the proposed tool set. Examples are: editing outlying observations with evolving data analysis tools, editing longitudinal data with time series techniques, editing at various levels of statististical units, or editing unofficial sources in order to assists official statistics. Such examples will be dealt with in subsequent development phases.

VI. CONCLUDING REMARKS

52 The proposed tool set is far from being the result of a complete and exhaustive evaluation by an independent authority. It rather corresponds to what was observed by several editing experts in the last few years of software development and applications for large production surveys. It is believed that the proposed tool set will offer coherent functions that can be run together one after the next in an integrated manner. It is also believed that it will enable the principles of the industrialization of editing as listed in Section II.B.

REFERENCES

Beaumont, J.-F. and Bissonnette, J. (2011). "Variance estimation under composite imputation: The methodology behind SEVANI". *Survey Methodology*, 37, 171-179.

Deguire, Y., Reedman, L. and Wenzowski, M. (2011). "Generalized Systems: The Statistics Canada Experience". *Proceedings of Statistics Canada Symposium*. November 1 – 4, 2011.

HLG-BAS. (2011a). "Strategic vision of the High-level group for strategic developments in business architecture in statistics". Document presented at the Fifty-ninth plenary session of the Conference of European Statisticians, UN-ECE, June 14-16, 2011.

HLG-BAS (2011b). "Draft Strategy of the High-Level Group for Strategic Directions in Business Architecture in Statistics". Document presented at the HLG-BAS workshop, Oct. 31 – Nov. 1, 2011.

ISTAT, CBS and SFSO (2007). "Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys". EDIMBUS. August 2007.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf

Poirier, C. (2011). "Industrialization of Editing – Statistics Canada Requirements", Internal Statistics Canada document. February 1st 2011.

Särndal, C.-E. (1992). "Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Statistical Network. (2010). “Informal CSTAT workgroup on stronger collaboration on statistical information management systems – Draft Output Report”. Document produced by the Value Creation Group, June 7 – 11, 2010.

Statistics Canada (2009). “Statistics Canada Quality Guidelines – Fifth Edition”. Statistics Canada Catalogue no. 12-539-XIE, October 2009.

Statistics Sweden (2011). “User´s Guide to SELEKT 1.1– A Generic Toolbox for Selective Data Editing”. Statistics Sweden document, February 17, 2011.

UNECE (2012). “Generic Statistical Information Model (GSIM): Overview – version V0.4”. Draft document under review.
http://www1.unece.org/stat/platform/download/attachments/65962906/GSIM+0.4+Overview.docx?version=1

Zhang, L. C., Brinkley, E., Coyle, J., Lodpell, J., Poirier, C., Svensson, J. and Zabala, P. (2011). “Industrialization of Editing: Business Case”. Working document produced within the Statistical Network.