CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Workshop 2019

26-28 June 2019, Geneva, Switzerland 14 June 2019

Challenges in re-designing the national metadata system according to international standards

Giorgia Simeoni, Mauro Scanu (Istat, Italy)

simeoni@istat.it, *scanu@istat.it*

Abstract

Istat has a long tradition in metadata systems. Concerning reference metadata and quality indicators of statistical processes, the SIDI/SIQual system (Information System on Quality of Statistical Processes) was developed in the 90s'. It is in production since 2001 and is currently updated. SIDI/SIQual model for the statistical business process documentation can be considered a precursor of the Generic Statistical Business Process Model (GSBPM), and the system also includes several quality indicators that have been adopted as standard by the European Statistical System. Furthermore, a system devoted to the description of the content and meaning of data sets (i.e. on structural metadata) has been under construction in Istat since 2010 under the acronym of SUM-MS (Sistema Unitario dei Metadati – Metadati Strutturali). The system took advantage of the availability of the Generic Statistical Information Model (GSIM) in order to identify its main metadata concepts. Nowadays, both the SIDI/SIQual and the SUM need to be re-designed as a new comprehensive national metadata system .

GSBPM and GSIM are considered as standard models supporting the system re-design. They should facilitate the identification of relevant concepts and information objects to be taken into account. Indeed, GSBPM seems to be useful to reorganise statistical process documentation already available in SIDI/SIQual; GSIM has already proved to be helpful in the design of SUM-MS as far as the Concepts and Structure Group concepts are concerned. However, the joint application of GSBPM and GSIM appears to be more challenging. Focusing on the information objects belonging to the GSIM Business Group leads immediately to terminology issues with GSBPM; at the same time, the interpretation and concrete implementation of some GSIM concepts is not always straightforward; finally, it should be assessed if GSIM and/or GSBPM are sufficiently detailed and structured to describe the complexity of SIDI/SIQual contents, given that it is not only a metadata system because it includes also quality metadata and standard quality indicators.

The paper will discuss the above mentioned issues, encountered in the design of the new system, as a case study of the joint application of GSBPM and GSIM standards, with the aim to stimulate discussion and trying to propose solutions that could also bring to improvements to the standard models themselves.

Keywords

National metadata system, GSIM, GSBPM

1. Introduction

The UNECE Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) are nowadays internationally recognised standards and are widely adopted in several National Statistical Institutes (NSIs) and International Organisations (IOs) around the world. GSBPM represents a common vocabulary to communicate on the statistical production process while GSIM supports and facilitates the modelling of (metadata driven) statistical information systems. Although they were developed as "complementary", with GSBPM describing the activities carried out in the statistical process and GSIM introducing the information objects necessary to describe inputs and outputs of such activities, the joint applications of the two models are less common and less straightforward. However, the need of better define the links between the two model has recently emerged in the international context. Indeed, one of the UNECE Supporting Standards Group tasks in 2019 is dedicated to Linking GSBPM-GSIM and experts from several NSIs and IOs are actively contributing to the task. Istat is among them and in the same time is working at NSI level to such a link, with the aim of using together the models in different applications. One of them is the design of a new comprehensive metadata system.

Indeed, Istat has a long tradition in metadata systems. The SIDI/SIQual system (Information System on Quality of Statistical Processes) on reference metadata and quality indicators of statistical processes was designed in the 90s' and implemented in the early 2000s'. The structural metadata system SUM-MS (Sistema Unitario dei Metadati – Metadati Strutturali) is being developed since 2010. SIDI/SIQual and SUM-MS are described in Section 2.1. and section 2.2. respectively. Both are currently used at Istat, but, for several reasons, it is now deemed necessary to re-design them, as a new comprehensive national metadata system. Such reasons together with the aims of the new metadata system are reported in section 3.

GSBPM and GSIM has been considered since the beginning as a reference for the new system design. Istat metadata experts considered as a precondition for the new system to be compliant with such relevant and widely adopted international standards and looked at them as they should support and facilitate the re-design. This revealed to be quite effective when applying the models separately, but less straightforward when trying to apply them jointly. The issues emerged and the solution identified are presented in section 4.1 for discussion. At the same time, the attempts to model the metadata already stored in the existing systems according to GSBPM and GSIM standards lead to the conclusions that some extensions/improvements of the models could be useful. Such proposals are presented in section 4.2.

2. Istat metadata systems

2.1 SIDI/SIQual

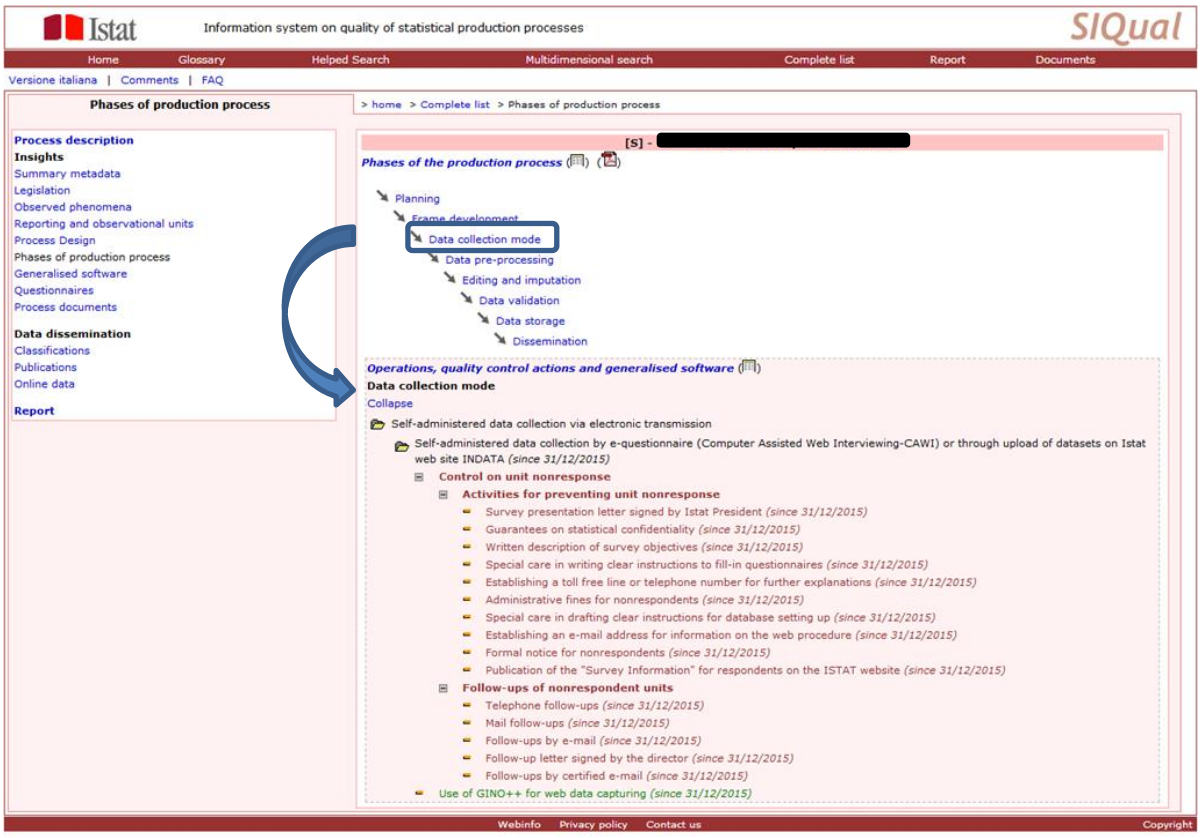
SIDI/SIQual is the official Istat system for documenting quality and reference metadata of statistical processes in terms of both qualitative documentation and quantitative quality indicators (Brancato et al. 2004, Brancato et al. 2006). The system is regularly used at Istat, and documents almost 400 statistical processes, 240 of which are currently being carried out, and the remaining suspended or ceased. Metadata and quality indicators update is scheduled on a mandatory basis once a year, although the existing net of quality pilots, in charge of documenting the processes, can update or change metadata in any time.

SIDI/SIQual is composed by two main subsystems: SIDI, the input system deployed over the Istat intranet, and SIQual, the navigation system, feeded by an ETL procedure. SIQual has two versions, one available on the Istat

web-site (http://siqual.istat.it), with the reference metadata, and one on the intranet, including also quality indicators.

In terms of contents, SIDI/SIQual can be considered a "quality oriented" metadata system. The core of the documentation is the description of *how* each statistical process is carried out. Such description is organised according to a model that has been mapped with GSBPM (Brancato and Simeoni, 2012). With respect to GSBPM, the Istat model is more structured and there is a particular emphasis on the quality control actions carried out to prevent, monitor and evaluate errors that can affect data quality. Also reference to generalised software used to carry out the different tasks can be reported (see fig. 1 for an example). In addition, the system includes several information on process objectives and contents, e.g. the topics observed and the units involved, links to press releases or other kind of dissemination means, and it is also a repository of relevant documentation (e.g. questionnaires, European regulation, paper and operative documentation).

Figure 1. Example from SIDI/SIQual corresponding to GSBPM "Run collection" sub-process



As already mentioned, SIDI/SIQual includes also several quality indicators. For example, a set of quality indicators related to unit nonresponse errors is collected for surveys (see fig. 2). One of the main characteristics of the SIDI/SIQual is that quality indicators are managed together with the metadata that are deemed relevant to interpret it (e.g. the reporting unit and the data collection mode for indicators on unit nonresponse errors). Most of the standard quality and performance indicators required in the European Statistical System (ESS) standard for reference metadata SIMS (Single Integrated Metadata Structure) are already included in SIDI/SIQual and others are planned to be included.

The richness of metadata and quality indicators managed in SIDI/SIQual made worthy the development of subsystems for quality reporting integrated to the system. They re-use the information stored in the system

database and re-organise it differently in order to satisfy different needs. Just to mention a couple of examples: a subsystem is devoted to produce the Methodological notes of the Italian Statistical Yearbook while another is able to produce the reference metadata files and quality reports in SDMX[1] format as required by Eurostat in the ESS Metadata Handler (Simeoni, 2013). Nevertheless, maintaining the "national" approach when international standards come out has its cost: it means mapping national concepts with international standard concepts and developing tools that translate as automatically as possible what is in the national system to be compliant with international standards.

Figure 2. Example from SIDI/SIQual: quality indicators on unit nonresponse errors

2.2 SUM-MS

A system devoted to the description of the content and meaning of data sets (i.e. on structural metadata) has been under construction in Istat since 2010 under the acronym of SUM-MS (Sistema Unitario dei Metadati – Metadati strutturali). The SUM-MS contains:

1. The definition of micro and macro data-structures produced along any statistical program undertaken by Istat.
2. The necessary concepts for micro and macro-data structures definitions.

Traditionally, at Istat these two kinds of metadata were not centrally managed. On the one hand, as in many other situations, there was a stove-pipe based management of metadata, where each pipe was a specific statistical program. On the other hand, metadata related to data produced in different process phases were also managed according to the storing and management facilities available for each specific phase, leading to a patchwork-like structural metadata management, actually unsuitable for any kind of industrialization process.

The SUM-MS was asked to modify this situation, centrally storing and managing definitions of data structures and the necessary concepts. Two main objectives were assigned to this new system: metadata harmonization between and within (along) statistical programs and traceability of each data production process.

[1] Standard for Data and Metadata eXchange

These two objectives were perfectly in line with the GSIM part on Concepts and Structures groups, whose definitions and components were immediately adopted in the SUM-MS. Although the nature of GSIM is to describe all the information objects available in a statistical program, including also those of the business part, the use of GSIM was restricted only to those concepts that refer to data (hence, Concept and Structure groups). The Business and Exchange parts were not used, yet.

3. New metadata system requirements

The existing Istat metadata systems SIDI/SIQual and SUM-MS are currently being re-designed as a new comprehensive national metadata system. SIDI/SIQual system needs to be renovated both to better document not traditional statistical processes like statistics based on multiple sources and for technological adaptation. SUM-MS should describe the content of data structures produced along the data production process as depicted by GSBPM: in this context, structural metadata on data structures content should be aligned and completed with process metadata.

The new system should be satisfy several demanding requirements from internal, national, and international users.

In order to satisfy Istat internal needs, the metadata system should be able to:

- preserve information already collected in existing metadata system SIDI/SIQual and SUM-MS, reorganising it if necessary;

- document, trace (and drive) different kinds of statistical production processes, from traditional surveys to the creation of multisource statistics;

- promote the use of harmonised metadata across the institute;

- reduce the "documentation" burden on production units obtaining information as a by-product of production processes, or through integration with other systems and re-using metadata in the system to satisfy external requests;

- manage and store the quality indicators that can be used as a basis to assess the quality of Istat statistical processes.

In order to satisfy national user needs, a metadata system should be able to provide metadata to accompany data dissemination, tailored for different kinds of users, from less to more expert ones. Such metadata should facilitate users in finding data they are searching, and support their interpretation.

Finally, to satisfy international requirements, in particular European Statistical System ones, a metadata system should be able to produce metadata according to ESS conceptual and technical standards, e.g. with Standard Code Lists, according to the Single integrated metadata Structure SIMS 2.0, translated in SDMX.

4. Use of UNECE Standards in the design of the new metadata system

GSBPM and GSIM are international standards that should support the new metadata system design. They should facilitate the identification of relevant concepts and information objects to be taken into account. Indeed, GSBPM seems to be useful to reorganise statistical process documentation already available in SIDI/SIQual, in a way that could be more comparable at international level; GSIM has already proved to be helpful in the design of SUM-MS as far as the Concepts and Structure Groups concepts are concerned. In order to model all the information

actually included in the existing SIDI/SIQual and the desiderata for the new system, also the Information Objects from GSIM Exchange and Business Groups should be taken into account. Although the design of the new system is still at the beginning, some issues in the use of GSBPM and GSIM appeared immediately. They are reported in the next sections for sharing experience and promote discussion and improvements in the models.

4.1 Joint use issue: terminology

One of the first step in the design of the new system was trying to identify correspondence between the concepts used in SIDI/SIQual and the standard models. Starting with GSBPM the task has been quite easy: general concepts in GSBPM are limited and the terminology used by GSBPM is quite similar with the one used ad Istat. Table 1 reports such correspondence. The detailed mapping between each specific SIDI/SIQual operation and GSBPM was prepared in 2012 (Brancato and Simeoni, 2012) and has been recently reported in SIDI/SIQual database. In the annex 2 there is an example of how the current documentation of a statistical process in SIDI/SIQual appears when reorganised according to GSBPM phases and sub-processes.

Table 1. Correspondence between GSBPM general concepts and Istat SIDI/SIQual elements

GSBPM concept	**Istat SIDI/SIQual element**
Statistical Business Process	Statistical production process
Phases	Phases of the production process
Sub-processes	Operations organised hierarchically (Istat operations are more detailed and describe how the subprocess is carried out)
Overarching processes	Operations or quality control actions

Similar task started to map SIDI/SIQual elements and corresponding GSBPM concepts with GSIM information objects and some difficulties suddenly came to light. The mapping with GSIM Information objects of the Business Group is needed to obtain a new integrated metadata system that allows not only documenting processes at reference metadata level but also tracing the data transformation process. Although GSBPM and GSIM have a common "origin", terminology used is quite different. Different terms are used for the same concepts and the same terms could have different meanings. For example the *GSBPM Statistical Business Process* correspond to the *GSIM Statistical Program*, while *GSIM Business Process* (seems to) correspond to *GSBPM Sub-process*. The latter match is not straightforward because in this way the concept of *GSBPM Phase* is missing in GSIM. A different solution could be to match *GSIM Business Process* with *GSBPM Phase* and *GSBPM sub-process* with *GSIM Process Step*. Indeed *GSIM Business Process* could not be split in further business processes, while this is possible for *GSIM Process Step*. In the international team working on the Linking GSBPM and GSIM the first option was adopted (i.e. *GSBPM sub-process* corresponding to *GSIM Business Process*), however, GSIM experts confirmed that the mapping could be done at different levels. At Istat we are still evaluating the different options. Unfortunately GSIM definition are often not clear: the *Business Process* is defined as "The set of *Process Steps* to perform one of more *Business Functions…"* and "A *Process Step* is a work package that performs a *Business Process*.", it seems a bit recursive. In addition, in the explanatory notes of *GSIM Business Function* it is reported that the set of GSBPM Phases and Sub-processes can be considered a high level catalogue of Business Functions. Thus should we consider that GSIM Business function correspond to GSBPM Phases and Sub-processes? In the end we decided to interpret GSIM Business function as the "purpose" that could be associated to each level (phase, business process, process step…) of the hierarchy. This seems to be possible reading the explanatory

notes, even if the Information Object is connected only to the Business process. However, this "room for interpretability" makes the model less useful.

Another issue discussed in the Linking GSBPM and GSIM task team was whether referring to the GSIM Information Objects related to the design or to the implementation layer in the linking exercise. Indeed, it is quite common in GSIM to have parallel Information Objects for the different layers. This is the case for example of the set of Information Objects devoted to the description of *Process Step*, *Process Input* and *Process Output* (with *Process Design*, *Process Input Specification* and *Process Output Specification* in the design layer correspondingly). For GSBPM users the reference to Design layer could unfortunately lead once again to terminology ambiguity referring to the activities carried out in the Design Phase. In the new Istat metadata system such duplication will be implemented only if strictly necessary to keep it as simple as possible. At the moment it seems that if a metadata system should drive the process execution on data the double layer is necessary, instead for documenting purposes the model could be simplified.

4.2 What is missing?

Variables, unit types and aggregate values

Both GSIM 1.1 and its new version 1.2 show the following definition and explanatory note of the concept "Variable":

"Definition: The use of a Concept as a characteristic of a Population intended to be measured.

Explanatory note: The Variable combines the meaning of a Concept with a Unit Type, to define the characteristic that is to be measured. Here are 3 examples: Sex of person, Number of employees, Value of production."

The definition and explanatory note seem to be contradictory. Sex cannot be considered as a characteristic of a population: it is a characteristic of a unit type, a single person. The GSIM explanatory note seems to be vague as well. It pinpoints the fact that the concept used as a variable must be associated to the generic conceptual counterpart of units in a population, i.e. a unit type. As said before, it is possible to recognize such association in the example "Sex of person" (sex is the variable and person is the unit type). In the same manner, does "Number of employees" decompose in a variable (number) and a unit type (employees, note the plural)? In Istat, the variable "Number of employees" is by itself a variable, attached to the unit type "enterprise".

If the concept "Number of employees" instead refers to the whole set of enterprises in the population of interest, it is an aggregate value, obtained conceptually by summing the number of employees of each enterprise in the population for all the enterprises in the population itself. In this case, the Neuchatel model on terminology allows the use of the concept Variable for "Number of employees": anyway, the Neuchatel model defines "Unit types" of two kinds: elementary and the aggregate unit types. Hence, number of employees can be considered as a variable, when associated to an aggregate unit type.

Our suggestion is that the definition of variable in GSIM should change accordingly, substituting Population with Unit type. Among Unit types, it should be stated the existence of elementary and aggregate unit types. After these changes, it becomes admissible that the term "Variable" can be attached to both individual and aggregate unit types:

- specifying that Sex of person (singular) is restricted to the elementary case,
- referring "Number of employees" to aggregate unit types, possibly grouped according to other variables (observed at the unit level) as the reference area or the NACE code of each enterprise.

Besides the terminology issues encountered trying to apply jointly and consistently GSBPM and GSIM to the design of the new metadata system, the mapping of main reference metadata concepts in SIDI/SIQual according to GSIM Information Objects is proceeding quite smoothly. In some areas it is clear that metadata already stored in SIDI/SIQual are more detailed or structured than in GSIM. For example the *GSIM Product* (Exchange group) could be further detailed similarly to how the different input sources are (i.e.: Questionnaire, Administrative register, Statistical register…). Documenting a paper publication could be quite different from a multidimensional table in a dissemination database.

Relevant reference metadata for a Statistical process are the legal requirements it satisfies (e.g. a European Regulation). These could be represented as the *GSIM Statistical Need* but it seems this concept could be detailed and refined.

Further area for possible improvements in both GSIM and GSBPM concerns quality metadata. As mentioned, SIDI/SIQual documents quality control actions carried out to prevent, monitor and evaluate errors that can affect data quality, and each action is formally connected to the operation whose quality it should assure. GSBPM Quality management overarching process includes such quality control actions, but there is not a connection with specific phases or sub-processes of the production process. Quality control actions could be modelled as specific GSIM Process Step but an Information Object or a Relation that allows relating such process step with the process step whose quality it tries to assure is missing. It is instead possible to model the generalised software used in the different business processes as *GSIM Business Service.*

Another issue is represented by the Information Objects of the Structure Group that should support modelling Reference Metadata Structure. The GSIM model in this case resembles the SDMX information model and it is quite simple. The Reference Metadata Structure corresponds to a template organised in a set of attributes and refers to a subject (e.g a survey). In GSIM a *Reference Metadata Attribute* is defined by a *Represented Variable.* This seems reductive. Indeed, the content of reference metadata attribute can be much more complex than a represented variable and it is very heterogeneous. The ESS standard structure for reference metadata is SIMS 2.0. it includes attributes that refer to policy adopted at NSI level as well as quality indicators that should be organised according to a data structure.

Finally, we are analysing if and how GSIM could facilitate modelling system improvements. SIDI/SIQual model for the documentation of the statistical process is not flexible in the routing among operations. The order of the workflow is predetermined with, e.g., editing and imputation always after data integration. The GSIM concept of *Process Control* is being taken into account to manage routing and improve system flexibility. At the moment it is not clear if it will be sufficient or other concepts will be needed.

5. Conclusions

As already mentioned, the design of the new Istat comprehensive metadata system is still at the beginning, but taking UNECE standard models GSBPM and GSIM as a reference supports such design.

In some areas we faced some difficulties in the application of the models. Hopefully same issues have been already faced and resolved by other NSIs that can share their experiences and findings.

In any case we tried to highlight areas where the standards could be improved or clarified as a further objective of the international community working on GSBPM and GSIM.

6. *References*

Brancato, G. and G. Simeoni. 2012. “Istat Statistical Process Modelling and the Generic Statistical Business Process Model: a Comparison.” European Conference on Quality in Official Statistics, Q2012, Athens, 29 May – 1 June, 2012.

Brancato G., Carbini R., Pellegrini C., Signore M., Simeoni G. (2006). “Assessing quality through the collection and analysis of standard quality indicators: the Istat experience”. The European Conference on Quality in Survey Statistics (Q2006), Cardiff, UK.

Brancato G., Pellegrini C., Simeoni G. e Signore M. (2004) “Standardising, Evaluating and Documenting quality: the implementation of Istat information system for survey documentation – SIDI”. Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004). Mainz, Germany.

Willeboordse et al (2006). Neuchatel model on terminology (PART II: Variables and related concepts object types and their attributes). https://statswiki.unece.org/pages/viewpage.action?pageId=14319930

Simeoni G. (2013), Implementing ESS standards for reference metadata and quality reporting at Istat, UNECE Work Session on Statistical Metadata, Geneva, 6-8 May 2013

Annex 1. Example of a SIDI/SIQual documentation for a statistical process reorganised according to GSBPM

GSBPM Phase	**GSBPM Sub-process**	**Istat SIDI/SIQual operation**
Specify needs	Identify concepts	Definition of the aims of the process
Design	Design outputs	Planning modes and technical devices for result dissemination
	Design collection	Planning data collection monitoring system
		Planning data collection modes
		Planning analysis unit contact and observation: direct interview, telephone interview, photocopy of administrative documents, etc.
		Questionnaire design
		Planning data entry modes
	Design frame and sample	Frame implementation planning
	Design processing and analysis	Planning editing and imputation methods
		Planning data analysis and tabulation procedures
		Classification and coding planning
		Definition procedures for automatic editing and imputation
Collect	Create frame and select sample	Development and updating of an Istat local frame
		Drawing the list of reporting units from an Istat centralised frame (e.g. Italian Statistical Business Register - Asia, Municipality Register, LAC)
	Run collection	Self-administered data collection by e-questionnaire (Computer Assisted Web Interviewing-CAWI) or through upload of datasets on Istat web site INDATA
	Finalise collection	Sorting out record sets of the same questionnaire
Process	Classify and code	Automatic coding (fully electronic)
	Review and validate	Manual error correction forcing wrong values on the base of experience
		Error detection based on consistency edits
		Manual error correction forcing wrong values on the base of deterministic rules as reported in a manual or document
		Manual error correction forcing wrong values on the base of previous correct data of the same survey
		Manual error correction forcing wrong values on the base of correct data from other surveys or sources
		Manual error correction based on reporting unit follow-up
	Edit and impute	Probabilistic error and outlier detection and nearest-neighbour donor imputation
		Deterministic error and outlier detection and imputation based on deterministic rules (IF-THEN)

		Deterministic error and outlier detection and nearest-neighbour donor imputation
Analyse	Validate outputs	Coherence control with previous data of the same survey
		Coherence control with data from other surveys or sources
		Control of influential unit data, strongly contributing to computation of aggregates (e.g. large enterprises)
		Coherence control among indicators of different analysis units of the same survey
Disseminate	Update output systems	Dissemination in Istat data bases accessible at the Data shop, Regional Offices or other locations
		Dissemination via Internet in an information system
	Produce dissemination products	Press release dissemination ("Statistiche-Flash", "Statistiche-Report", "Statistiche-Focus", "Note informative")
		Publication of final data on survey-specific volumes (Series Yearbooks, "Informazioni", etc.)
	Manage release of dissemination products	Access to the laboratory for the Analysis of Elementary Data (ADELE)