Implementing the quality framework for the Istat Integrated System of Statistical Registers: challenges and solutions[1]

Cecilia Casagrande, Sara Giavante, Fabiana Rocci and Giorgia Simeoni[1]

[1]Italian National Institute of Statistics (Istat), Italy

Abstract

One of the pillars of Istat modernisation programme, started in 2016, is the creation of the Integrated System of Statistical Registers (ISSR). Each register of the ISSR is the result of the integration of several administrative data sources and possibly survey results. Thus, the processes underlying the statistical registers are very complex due to their multisource nature and the need for coherence, within and across the registers, of the produced results. To monitor such a complex system, Istat developed a new quality framework, based on a metadata model that refers to the UNECE standard GSBPM (Generic Statistical Business Process Model) and including several quality measures. The framework assures a structured and detailed documentation for transparency and traceability reasons and allows assessing processes and outputs quality, both while they are in progress and ex-post, in a systematic and standardised way. For each GSBPM sub-process considered relevant for the statistical registers' processes, the set of the possible input, statistical methods and outputs is specified, as well as a set of standards quality indicators for monitoring and evaluating purposes. The metadata are needed not only for documentation purposes but also to provide the information useful to calculate and properly interpret the quality indicators. The framework was tested on two statistical registers of the ISSR, confirming its validity and usefulness but also highlighting the need for a certain degree of customisation when applied in each register. The implementation started in four statistical registers through different working groups in a parallel way: for each of them the processes should be first mapped with the GSBPM, then metadata should be compiled and the applicability of quality indicators should be evaluated. Sometimes quality indicators have to be tailored to the register to make them meaningful and useful. An informal restricted group of expert of the framework, involved in the different applications, is sharing those experiences, in order to deal with issues or doubts that may arise, as well as to assess ideas for possible improvements of the framework itself. In this way, the coordination and coherence between the implementations is guaranteed through discussion and problem-solving analysis. The paper will describe briefly the framework, the further challenges that the coordination group is encountering, the solutions identified and how the achievement of the final fine-tune of the quality framework is planned.

Keywords: statistical registers, quality monitoring, quality indicators, metadata

1. Introduction

Many National Statistical Institutes have invested in recent years to create a coordinated system of statistical registers capable of leveraging data from various sources, improving the quality of collected information, reducing data collection costs, and minimizing respondent burden.

The Italian National Institute of Statistics (Istat) began this renewal process in 2016 by creating the Integrated System of Statistical Registers (ISSR) aimed at maximizing the yield of

[1] This paper resumes the outcomes of a work jointly carried out by the authors, however Sections 1 and 4.2 are attributable to Sara Giavante, Sections 2 and 5 to Fabiana Rocci, Sections 3 and 4.3 to Giorgia Simeoni, Sections 4 and 4.1 to Cecilia Casagrande. Section 2.1 was jointly written.

information from administrative data combined with survey data to produce official statistics. Further details on this system can be found in section 2.

Evaluating the quality of such a system required an investment in methodological research to formalize a new approach for the use of administrative data combined with data from other sources systematised in statistical registers. Therefore, Istat has structured a quality framework that captures these specificities, starting from international standards (such as GSBPM[2] and GSIM[3]) and adapting such existing theoretical models to a context based on registers. The framework described in section 3 includes metadata and quality indicators related to the phases of the statistical process, appropriately divided into sub-processes, and is organized into metadata sections to capture the peculiarities of each sub-process as effectively as possible. Such theoretical model is now in the implementation phase across four statistical registers different in terms of nature, purpose and stage of implementation. The registers involved are: the *Base Register of individuals*, the *Thematic Register of Labour,* the *Thematic Register of Education and Training*, and the *Extended Register of Public Administration Units*. These registers will be briefly described in section 2.1. The focus will be then posed on those elements that are emerging during implementation and require coordination and continuous dialogue among people involved in the project, in order to improve the framework itself and facilitate next applications. Indeed, some issues prompted discussion and are requiring an adaptation in the implementation strategy or a refinement of metadata and quality indicators proposed, in order to make the framework sensitive and capable of capturing the various facets that emerge during the quality analysis of the different registers (Section 4).

2. The Italian Integrated System of Statistical Registers

During the last decade, Istat has been engaged in a modernization program involving the revision of the statistical production model. The modernization of the production processes has been achieved by preparing the Italian ISSR. The ISSR is a complex system that get started by the integration of one or more different type of sources (administrative registers, surveys and other statistical registers). It was mainly built in order to support the consistency of statistical production processes and to improve the quality of information for users substituting the “silos” model adopted before.

---

[2] GSBPM: Generic Statistical Business Process Model, for further information https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51

[3] GSIM: Generic Statistical Information Model, for further information https://unece.org/statistics/modernstats/gsim

The ISSR is composed of different types of registers, whose definitions can be based on different combination of the following elements:

a. the type of target information the register is designed for;
b. the features of the statistical units on which it is based.

The first distinction that arises is between Base and Satellite Registers, which are in turn divided between Extended and Thematic Registers The following definitions apply:

- Base Statistical Registers (RSB): the target information is what is necessary to recognize a unit as belonging to a specific population of official statistics. In this context, Istat examples of RSBs are (Istat, 2016c): i) Base Register of Individuals; ii) Register of economic units, enterprises and institutions (ASIA); iii) Base Statistical Register of Places.
- Satellite Statistical Registers: their purpose is to release variables for specific thematic phenomena, for example, education, health, safety, income, etc. Their subsequent subdivision depends on the type of statistical unit underlying them, so we have:

  - Extended Statistical Registers: the statistical units can be identified among those belonging to one of the base registers while the variables concern i specific phenomena, often identified by EU regulations.

  - Thematic Statistical Registers: these registers also provide information on specific phenomena but differ from the extended ones because the fundamental statistical unit is specifically created, usually linking several base statistical units through a relation typical of the phenomenon under study.

In the next subsection the four registers of the ISSR to which the quality framework is being applied will be briefly introduced.

2.1 Statistical registers of interest

The *Base Register of Individuals* represents the reference register to produce official statistics regarding the population. The process of constructing the register involves the integration of more than 50 different administrative data sources, beside demographic data and results of social surveys. The register contains core variables that remain unchanged over time, such as gender, date and place of birth, and core variables that may change over time, such as citizenship, level of education, and marital status. The variables of the register follow diversified construction processes due to their peculiarities, often linked to the release timing of the involved sources. Therefore, this register proves to be complex in its implementation not only due to the multitude of involved sources but also because of their diversified nature.

The *Extended Register for Public Administrations Units* contains structural and economic variables on a subset of the Italian Public Administrations (PA). The statistical units belong to

the part of ASIA related to the PA, the input data are the official balance sheet of the PA Institutions, both for income and expenditures.

The *Thematic Register of Labour* represents the reference Italian framework of data and metadata for the estimates of employment, wages and compensation of employees. The base unit is “Job”, a concept that identifies the basic element of the register. Job is defined as “a relationship between an economic unit and a person having as its object a work activity”. The main variables considered into this register are useful to produce statistical information about all labour topics and they include different employment measures as labour inputs, labour cost factors and labour incomes. The sources of information that are involved to build this register are mainly administrative. Those sources are analysed, organized and integrated in order to achieve a statistical set of information useful also to serve the purposes of other several statistical production processes.

The *Thematic Register of Education and Training* is aimed to provide official yearly statistics on education and training related to individual (e.g. education level) as well as to education institutions (e.g.: schools, universities). The register is mainly based on the integration of administrative data from different sources. The core unit is the “education position”, identified by the combination of three elements: the individual, the institution and the education and training program. This register is still in the design phase and its release is planned for 2026.

3. The Quality framework for the ISSR - QSIR

In 2019, while the development of the ISSR was still ongoing, Istat started working also to build the quality assurance layer to be integrated in the ISSR processes. It should allow their documentation, quality monitoring and evaluation.

Through internal working groups including thematic, methodological, metadata and quality experts, a comprehensive framework, named QSIR, has been defined (Di Zio et al. 2023). The QSIR proposes standard metadata items and quality indicators to be applied to the different steps of the statistical register processes. To identify such metadata elements and quality indicators, a thorough analysis of the processes underlying the creation of the registers of the ISSR was first carried out. The analysis led to the identification of the most relevant steps, that have then been also mapped on the Generic Statistical Business Process Model – GSBPM (UNECE, 2019) as reported in Table 1.

Since the framework should be applied to the current editions (cycles) of each register and not to the initial design and implementation, during which ad-hoc quality evaluations were carried out, most of the GSBPM sub-processes of the first 3 phases were not considered. Indeed, only

the “Check data availability” sub-process was considered, to take into account the possibility of variations in the availability of the sources in different editions. In addition, the modernisation process at Istat brought to the centralisation of data collection activities and now the Directorate of Data Collection not only follows centrally the acquisition of data from surveys and administrative sources, but it carries also out the first technical checks on the data and the pseudonymisation process. Consequently, these phases are not considered in the QSIR framework.

Table 1: Relevant sub-processes for ISSR mapped with GSBPM

QSIR Sub-processes	GSBPM corresponding Sub-processes
Check data availability	1.4 Check data availability
Acquire data	4.3 Run Collection
Conduct preliminary evaluation	8.2 Conduct evaluation
Integrate data	5.1 Integrate data
Classify and code	5.2 Classify and code
Edit and impute	5.3 Review and validate, 5.4 Edit and impute
Derive new variables and units	5.4 Derive new variables and units
Calculate aggregates	5.5 Calculate weights; 5.6 Calculate aggregates
Validate outputs	6.2 Validate outputs

The most peculiar sub-processes in the ISSR context are certainly the ones related to the preliminary evaluation of the dataset and to the data integration. In the former, activities like deduplication, checks on missing data and, when possible, an evaluation of the coverage of the data with respect to the target population is included. In the latter, the different methodologies for combining different sources are considered. The two sub-processes are described in Appendix 1. For each of the relevant sub-processes a set of quality indicators was designed specifically for monitoring and quality assessment purposes: in Appendix 2 the lists of quality indicators defined for the sub-processes *Conduct preliminary evaluation* and *Integrate data* are reported as examples.

At the same time, for documentation, transparency and traceability purposes, but also to be able to automatise the calculation of the quality indicators as well as to allow their correct interpretation, proper metadata describing possible inputs, outputs and statistical methods applied in each sub-process have been clearly identified, on the basis of the model developed by UNECE (UNECE, 2022) that links the standards GSBPM with GSIM (Generic Statistical Information Model, UNECE, 2019b). In the metadata templates in Appendix 2 (first two columns) you can see the different GSIM Information Objects considered. In the third column the set of possible values that each metadata element can assume have been identified for

the *Conduct preliminary evaluation* and *Integrate data* sub-processes. Applying the framework QSIR in a statistical register means describing the process through the metadata templates, defining the workflow that links the sub-processes, identifying which of quality indicators proposed is applicable and meaningful and test it. All this work will be the input for the IT sector to develop, on the one hand, automatic procedures in the registers monitoring systems that automatically calculate the indicators and make them available to thematic experts working for their daily monitoring activities; on the other hand, the new metadata system Istat is currently designing, METAstat, that will store both metadata and quality indicators to document and evaluate all Istat statistical processes (not only statistical registers).

4. Implementation issues

As already mentioned the QSIR framework is currently being implemented in 4 different registers, that are of different types (base, extended, thematic) and also at different stages of development and of standardisation level in the architecture. Even if the QSIR was tested before its release, during the actual implementation phase several new issues are arising that need to be addressed: they can also drive to a refinement of the framework or to develop better strategies for future implementations. In order to address these challenges, experts of the framework decided to set an informal group to discuss the issues and share solutions. Here after a few examples of the issues that should be faced and the solutions identified are reported.

4.1 Check data availability metadata template

The “check data availability” template has been defined in order to monitor every change that could happen on the usual sources expected and needed for the register under construction. The QSIR has been released with the following typologies of variation:

- A usual source is no more available;
- A new source is being evaluated by analysing its fitness for statistical purposes (in terms of punctuality, safety and stability of the data provision);
- Unexpected changes are present in a usual source and it is necessary to control whether the quality requirements of the source are still maintained.

The activity conducted on different registers allowed to observe how an exhaustive mapping of each source of information could be very important to monitor the content of the sources and their main characteristics in terms of metadata available, instead of only consider the changes in the sources.

Moreover, it the opportunity to add a new indicator not considered yet for this step has been evaluated. It is “the sources presence/absence indicator”, to verify at the beginning of the production process whether the sources are all available or not, in order to discover other problems than the ones listed above (for example, an important delay in the supply of data). In the first case, the process enables to go ahead to the next step, otherwise it is suggested to evaluate what is the impact of what is not available by calculating the correct indicators listed. In this way, the template is not only a descriptive tool, but it also provides guidance on the available sources and to the next steps according to the actual scenario of data, fulfilling its monitoring and improving quality purposes.

4.2 Customisation of quality indicators

Although the QSIR framework aimed at defining a standardised set of quality indicators that could also allow harmonised quality evaluations and comparisons between different registers, it has always been clear that the QSIR quality indicators could not be exhaustive for the purpose of monitoring a specific register. It was well expected that additional specific quality indicators could be needed for each register. What was not expected was that to apply the QSIR quality indicators in different registers specific customisation could be needed. A simple example is the *link rate* that should be calculated in case of data integration. Its implementation depends on the integration strategy adopted in the register. If there is a prioritisation of the different sources to produce the register, the quota of units of each source that is linked with in the integrated dataset could be not representative of the integration process and not useful for monitoring purposes, while could be useful to know the origin of the units of the integrated dataset. It was thus decided to allow such a customisation to obtain more meaningful and useful indicators in monitoring systems. This implies that accompanying metadata should clarify the customisation: this will be necessary in particular when the quality indicators will be collected centrally for documentation purposes by METAstat, in order to avoid wrong comparisons and conclusions.

4.3 Adaptive implementation strategy

The different registers on which the QSIR framework is being applied are at different stages of development and, even if the architecture of ISSR is harmonised to a great extent, there are some differences in the implementation approaches in different registers. In some cases, like for the *Thematic Registers of Labour* and the *Thematic Register of Education and Training*, a template for documenting the different steps of the register process is already in place. Such template was developed mainly by IT experts that developed the software components that implement the sub-processes. Sometimes there is correspondence between such components

and the QSIR sub-processes, sometimes the granularity is different (e.g. one software component implements only a part of a sub-process, or implements more than one sub-process). In any case, such documentation was already available, familiar for register and IT experts, and included several information useful for QSIR templates compilation, like core input, output and software. Thus, as a strategy to facilitate implementation of the QSIR framework, it was decided to re-use it, map with QSIR template and ask to register experts only to integrate the missing information (e.g the description of the process method and the identification of quality indicators that could be used).

5. Concluding remarks

Official statistical production is increasingly relying on multisource statistical processes, that are much more articulated than traditional surveys. In each register of the Italian ISSR different variables can be originated by different sources applying different workflows and methodologies. This makes difficult to set up monitoring systems. The QSIR framework aims at facilitating this task, while providing also structured and standard solutions for documentation and ex-post quality evaluation of such processes. In this paper the framework was presented, but the focus has been voluntarily posed more on issues arisen during implementation and how they have been solved than on the framework itself. The purpose was to share the successful experience on the implementation of a demanding new approach for documentation and quality assessment, achieved through discussion, collaboration spirit and flexibility. It has also been highlighted that such approach sometimes also led to adjustments and refinements of the framework itself. Finally, another objective that is being persecuted during implementation to reduce the burden of its application during current production is to automatise as far as possible. At the moment the calculation of quality indicators is being automatise, but next step could be the definition of a system of quality gates, e.g. thresholds for the indicators values that produce warnings only in case of out-of-control values, to further facilitate the monitoring task of staff working in registers management.

References

Di Zio, M., Falorsi, S., Rocci, F., Simeoni, G. (2023). Process and output quality evaluation measures for Istat Integrated System of Statistical Registers *EESW 23*, Lisbon, 20-22 September 2023.

HLG-MOS, UNECE. (2023). The Generic Statistical Information Model (GSIM v. 2.0), available on-line: https://statswiki.unece.org/display/gsim (retrieved on 13/07/2023).

Istat. (2016). Il Programma di Modernizzazione dell'Istat https://www.istat.it/it/files/2010/12/Programma_modernizzazione_Istat2016.pdf.

Istat. (2023). Monitoring and Evaluating the Quality of the Integrated System of Registers, Istat *Working Papers N. 7/2023*.

Appendix 1 Metadata templates examples

Table 1. Metadata template for "Conduct preliminary evaluation“

In this sub-process preliminary checks and evaluations are carried out on a dataset, which can be coming from individual sources or be an integrated dataset. Deduplication is performed and not usable records are identified and deleted. Missing values are checked. If an appropriate auxiliary source is available (benchmark), the coverage of the dataset under consideration can be estimated, both on the basis of aggregate comparisons and through micro-level matching. The coverage estimate is accompanied by the evaluation of the representativeness of the dataset with respect not only to the number but also to the characteristics of the units contained in it, for example: presence or absence of large companies, of universities with unique or rare degree courses.

The indicators calculated in this phase on the individual sources may be a useful feedback for the unit identification phase, in particular those relating to deduplication, given that these errors may be specific to the source or due to the pseudonym attribution process,

Macro Item	GSIM Object	Possible values
Input	Core input	Data-set to be evaluated (data structure: units and variables): it can be one of the source dataset or the integrated data set
	Parameter input	Key linkage variables
	Process support input	Reference/benchmark data-set Definition of the population of the data-set to be evaluated
GSBPM sub-process	Business Function	Deduplicating data-set, checking for missing values, evaluating the coverage of the dataset
	Business Process (GSBPM phase)	8. Evaluate
	Process Step (GSBPM sub-process)	8.2 Conduct evaluation
	Process Method	Duplications identification Identification of missing or not usable data Coverage evaluation through aggregate comparisons Coverage evaluation through microdata matching
	Rule	Rules to identify the deduplication Rules to identify missing values Rules to identify not usable data Integration model, relationship 1-1, n-1, n-n
	Software Agent	Relais, Statmatch, Ad hoc procedures
Output	Core output	Data set without duplicates or unusable data
	Process Metric (Quality indicators)	See appendix 2
	Process Execution Log	Processing time

Table 2. Metadata template for "Integrate data" sub-process

In this sub-process the integration activity between different sources takes place with the main objective of building a largest and most complete dataset possible of the variables of interest in the register itself. Integration can take place at different moments of the statistical production process (immediately after data collection or downstream of processing individual sources); it can involve all sources simultaneously or subsets of sources sequentially and therefore can be repeated several times within a statistical production process. Possible integration objectives are:

- to increase the units i.e. improve the general coverage of the data (horizontal integration at the level micro);

- to increase the variables i.e. increase the availability of data (vertical integration at the micro level);

- to increase units and variables (horizontal and vertical integration at the micro level);

- to construct new units or new variables not present as such in the individual sources but obtainable from functions involving data from different sources.

Integration can also take place after the construction of a register, to integrate it with another

ISSR register, for example for validation purposes.

Macro Item	GSIM Object	Possible values
Input	Core input	Data-set1, Data-set2, … (data structure: units and variables)
	Parameter input	Threshold, Linkage keys, Blocking variables
	Process support input	Further variables useful for identification other than the keys or to control the matching
GSBPM sub-process	Business Function	Increasing units, increasing variables, increasing both
	Business Process (GSBPM phase)	5. Process
	Process Step (GSBPM sub-process)	5.1. Integrate data
	Process Method	Record linkage (deterministic, hierarchical, probabilistic, privacy preserving and predictive linkages (classification or regression techniques) Statistical matching Appending procedures Data pooling Integration based on data source prioritisation
	Rule	Integration model, Rules for the hierarchical selection of the sources, transformation rules
	Software Agent	Relais, Statmatch, Ad hoc procedures
Output	Core output	Integrated Data set, Non linked records data sets
	Process Metric (Quality indicators)	See appendix 1
	Process Execution Log	Integration time

Appendix 2 Examples of quality indicators

Box 1. Quality indicators for “Conduct preliminary evaluation” sub-process

Deduplication indicators
3.1. Percentage of duplicates records
3.2. Percentage of duplicates records on the key variable
3.3. Percentage of duplicates records on a set of relevant variables
3.4 Discrepancies between information present in duplicate records
Missing values indicators
3.5 Missing value rate for the main variables
3.6 Percentage of not usable records
Coverage and representativeness indicators
3.7 Coverage rate of the evaluated dataset with respect to the benchmark dataset
3.8 Comparison between statistics (average, totals, …) and distribution of variables between the evaluated dataset and the related sub-population in the benchmark dataset
3.9 Comparison between statistics (average, totals, …) and distribution of variables between the evaluated dataset and the total population in the benchmark dataset

Box 2. Quality indicators for sub-process “Integrate data” sub-process

4.1. Missing values or errors in linkage variable
4.2. Match rate
4.3. False link rate
4.4. False non-link rate
Indicators on units
4.5. Percentage of units from different datasets on unit total
4.6 Under-coverage of administrative dataset
4.7 Over-coverage of administrative dataset
Indicators on variables
4.8 Percentage of variables from different input datasets on total number of variables in the integrated dataset
4.9 Distances between variable distributions on the integrated dataset and on the input datasets
4.10 Number of variables derived at the end of integration
4.11 Incoherence in the information present in the different sources on linked records