Integrating international standards in the design of reference metadata component of the new Istat system METAstat[1]

Francesca Budano, Giorgia Simeoni[1]

[1]Istat, Italy

Abstract

Since the beginning of 2000s, Istat has been equipped with a rich and a well-structured quality and reference metadata system, SIDI-SIQual. It is aimed at documenting and supporting quality monitoring and assessment of statistical production processes. SIDI-SIQual describes the production process and its features: information content; survey phases and operations; activities to prevent, monitor and evaluate sampling and non-sampling errors, standard quality indicators that are both process and product-oriented, etc. Since the system is well tailored for traditional surveys but less effective for innovative, e.g. multisource or experimental, statistical processes, and also technically obsolete, Istat is designing a new metadata system, called METAstat. The new system is going to manage both structural and reference metadata in an integrated way. With regard to the process documentation, METAstat model relies on well-known international standards like the UNECE Generic Statistical Business Process Model – GSBPM, and the Generic Statistical Information Model – GSIM. The two models are used complementarily to describe the phases and sub-processes of each type of statistical process in terms of inputs, outputs and methods used. Thanks to the flexibility of these models METAstat overcomes the limitation of current system SIDI-SIQual, allowing to describe innovative and complex processes, e.g. detailing different sources and data processing steps to produce different variables or modules in the same process or specifying innovative data and methods used for producing experimental statistics. In addition, the quality layer of statistical processes, that in GSBPM is described as an overarching process, in METAstat is detailed and integrated in each sub-process, including standard quality indicators, similarly to the current SIDI-SIQual approach. This allows to derive from the statistical process documentation almost all the information needed to produce quality reports according to the European Statistical System Standard SIMS (Single Integrated Metadata Structure). Finally, METAstat is going to be integrated with many other Istat information systems, from the one managing the National Statistical Programme to the repository of validated microdata, in order to collect metadata and quality indicators as much as possible in an automatic way. In this way the implementation of METAstat will reduce the documentation burden on production units and at the same time improve coherence and comparability of metadata and quality indicators across different statistical processes. The paper will describe the model developed for reference metadata and quality documentation in METAstat with a particular focus on how to integrate GSBPM, GSIM and SIMS.

Keywords: metadata system, reference metadata, GSBPM, SIMS

[1] This paper resumes the outcomes of a work jointly carried out by the authors, however Section 1,2,3 are attributable to Francesca Budano while section 4 and 5 to Giorgia Simeoni.

1. Introduction

Istat can be defined a "precursor” in the documentation of statistical processes and their quality: in the beginning of 2000s, the SIDI-SIQual system was developed as the Official Istat system for documenting quality and reference metadata of the statistical production processes, with the aim of monitoring the quality and performance of statistical processes.

Nowadays SIDI-SIQual system presents some weaknesses: first, itis obsolete from an IT point of view and, second, it is not enough flexible to well describe innovative statistics based on multisource processes, statistical registers or statistical processes based on new sources data (e.g. big data).

For this reason, Istat started a project on design and development of a new metadata system called METAstat; itis going to manage and integrate reference metadata, structural metadata and terminology and will allow to obtain more flexibility for documentation, and a greater level of detail for traceability and will facilitate quality monitoring and assessment of statistical processes while promoting metadata harmonisation. It will be based on the integrated use of international standard models and will be aimed at reducing the burden on production sectors in Istat.

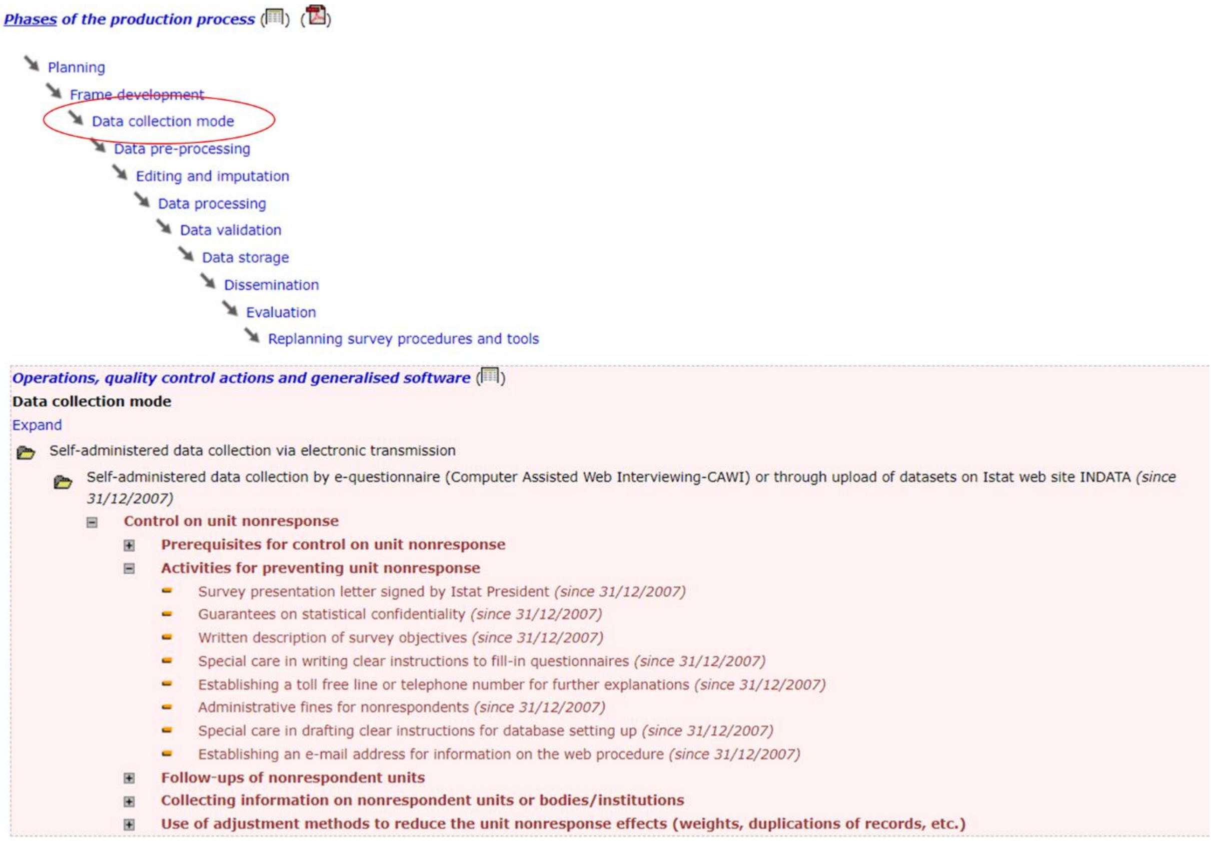
This paper focus on the reference metadata component of METAStat and how the documentation modelled according to GSBPM (Generic Statistical Business Process Model, UNECE, 2019), and GSIM (Generic Statistical Information Model, UNECE, 2023) standards will allow to produce automatically the quality reports according to European Statistical System (ESS) standards SIMS (Standard Integrated Metadata Structure, Eurostat, 2021).

2. SIDI-SIQual

The SIDI-SIQual system (Brancato et al. 2006) is accessible to external users through SIQual (https://siqual.istat.it) and describes the Istat statistical production processes and their features, with particular reference to quality issues: indeed, survey phases and subprocesses are supplemented with actions to prevent, monitor and evaluate sampling and non-sampling errors, .

Figure 1 illustrates an example of documentation of the data collection phase for a survey that use a CAWI (Computer assisted web interview) mode in SIDI-SIQual in terms of metadata and quality control actions.

Figure 1. The data collection subprocess of an Istat survey



Phases of the production process

- Planning
  - Frame development
    - Data collection mode
      - Data pre-processing
        - Editing and imputation
          - Data processing
            - Data validation
              - Data storage
                - Dissemination
                  - Evaluation
                    - Replanning survey procedures and tools

Operations, quality control actions and generalised software

Data collection mode

Expand

- Self-administered data collection via electronic transmission
  - Self-administered data collection by e-questionnaire (Computer Assisted Web Interviewing-CAWI) or through upload of datasets on Istat web site INDATA *(since 31/12/2007)*
    - ⊟ **Control on unit nonresponse**
      - ⊞ **Prerequisites for control on unit nonresponse**
      - ⊟ **Activities for preventing unit nonresponse**
        - Survey presentation letter signed by Istat President *(since 31/12/2007)*
        - Guarantees on statistical confidentiality *(since 31/12/2007)*
        - Written description of survey objectives *(since 31/12/2007)*
        - Special care in writing clear instructions to fill-in questionnaires *(since 31/12/2007)*
        - Establishing a toll free line or telephone number for further explanations *(since 31/12/2007)*
        - Administrative fines for nonrespondents *(since 31/12/2007)*
        - Special care in drafting clear instructions for database setting up *(since 31/12/2007)*
        - Establishing an e-mail address for information on the web procedure *(since 31/12/2007)*
      - ⊞ **Follow-ups of nonrespondent units**
      - ⊞ **Collecting information on nonrespondent units or bodies/institutions**
      - ⊞ **Use of adjustment methods to reduce the unit nonresponse effects (weights, duplications of records, etc.)**

For the data collection mode, a lot of quality control actions can be specified: in table 1 actions related to control on unit nonresponse are reported. SIDI-SIQual documents the statistical process phases with quality control actions but it is not flexible: indeed, it is not possible to specify different actions or workflows for different variables or specify inputs and outputs. Istat internal users of the system can find metadata accompanied by several process and product oriented quality indicators. Such indicators are useful for quality assessment of specific processes, while the analysis of long time series of quality indicators can also allow to evaluate the impact of important innovations introduced in the process or the effect of unexpected events such as the COVID-19 pandemic.

3. The role of international standard models

METAStat is going to manage both structural and reference metadata and will rely on most relevant international standards related to quality and metadata (e.g. GSBPM, GSIM) to overcome the limitations of the current SIDI-SIQual system.

In particular, GSBPM will be used in METAstat to describe statistics production processes of any type: from traditional surveys, to Trusted Smart Statistics including statistics based on administrative data and multisource processes .

GSIM is a standard model developed by UNECE to represent statistical information objects. It provides a common framework and will used in METAstat to describe concepts, processes and statistical data.

GSBPM and GSIM are complementary to describe phases and sub-processes of each type of statistical process in terms of inputs, outputs and methods used (UNECE, 2022).

GSBPM and GSIM are used as a reference to design the conceptual model of METAstat, but the system aims to be the unique Istat metadata system, through which also standard quality reports should be produced. SIMS is a template that facilitates the harmonized and efficient preparation of quality reports, both producer- and user-oriented. It Includes conceptual, methodological and quality metadata, as well as quality and performance indicators. It is the standard for the reference metadata and quality reports in the ESS and it is also the object of a recent recommendation approved by the European Commission (Commission recommendation (EU) 2023/397 of 17 February 2023).

SIMS and GSBPM are both models to document statistical processes and statistical outputs, but SIMS is more oriented to output quality aspects, while GSBPM is more oriented to process description. In appendix 1 the main figures representing GSBPM and SIMS are reported as a reference.

4. METAstat

METAstat aims to be the new unique Istat metadata system integrating structural and reference metadata as well as terminological issues (i.e.: the glossary). It will be a repository of harmonised metadata providing useful services to the Institute. It will allow to fulfill transparency and traceability requirements, facilitate the re-use of harmonised metadata and support quality monitoring and assessment procedures.

It will be integrated with several other information systems at Istat in order on the one side, as input, to automatise as far as possible the collection of metadata and quality indicators and consequently reduce the documentation burden, and, on the other side, to provide harmonised metadata to other systems for many different purposes.

With regard to reference metadata, METAstat should overcome the weaknesses of current SIDI-SIQual and provide more detailed and re-usable documentation. The plan is indeed to document a statistical process once and to produce automatically from such documentation different outputs according to the objectives: e.g. standard quality reports to be transmitted to Eurostat or published in Istat website as well as Methodological documents to accompany microdata release, or Short metadata notes to be included in general publications as information on the data source. In order to obtain these goals the reference metadata component is being designed with a modular structure. The following modules will be progressively implemented:

1. General information and main informative contents of a statistical process

2. Statistical process inputs and outputs

3. Statistical process phases and subprocesses (including quality indicators)

The conceptual model of each module is based on GSIM and GSBPM, but enriched with information that proved to be useful in the SIDI-SIQual experience. As already mentioned, the information collected will be then re-used to automatically fill in the quality reports according to SIMS. A similar functionality is already implemented in SIDI-SIQual (Simeoni, 2013), but not all the SIMS concepts could be obtained automatically, further editing was needed and the integration made in the SIMS texts were not reported in SIDI-SIQual documentation.

In the following subsections the design of the different modules will be described and the corresponding SIMS concepts that will be consequently filled in will be pointed out.

4.1 General information and main informative contents of a statistical process

The main object in METAstat is the statistical process. METAstat statistical process corresponds with the GSBPM Statistical business process and the GSIM *Statistical Programme*. As almost every Information Object (IO) in GSIM[2] among the *Statistical Programme* attributes there are the ID, the name and the description. Then, as specific attributes, we can find from when the *Statistical Programme* has started and the legal framework (e.g., legal basis for the statistics to be produced by *Statistical Programme*). In addition, according to GSIM different *Agents* can be connected with the *Statistical Programme* with different *Roles*. For example we can have the *Individual* responsible for the *Statistical Programme*. Such *Agent* would have also a function, an organisational structure and some contact information.

[2] To be precise it is true for each information object that is a sub-type of Identifiable artifact

These GSIM objects are represented in Appendix 2, figure 3, where it is also shown how the information perfectly corresponds with specific concepts required by SIMS. In particular:

- the description of the *Statistical Programme* can be reported in SIMS S.3.1 Data description;

- the period from which the *Statistical Programme* has started can be the reported as the initial period in S.3.8 Time coverage;

- the legal framework is the input for the S.6.1 Legal acts and other agreements ;

- and the *Agent* characteristics can be reported in the S.1 Contact sub-sections.

Further metadata like the Information needs satisfied by the *Statistical Programme*, the target population and the statistical units observed are modelled with corresponding GSIM objects and can be easily mapped with SIMS concepts.

4.2 Inputs and outputs of the statistical process

Moving from general information on the statistical process to the description of the statistical process edition (or *Statistical Programme Cycle*, according to GSIM), in designing METAstat it has been considered useful to distinguish external inputs and final outputs of the whole statistical process from intermediate inputs and outputs of specific phases and subprocesses. In figure 4 of Appendix 2 different types of external inputs like administrative data or statistical registers (*Register*) or *Classifications* are shown, as well as different kinds of final outputs (*Product*): Corporate dissemination database, News release, Publications or microdata. Also in this case the SIMS correspondence (S.18.1 Source data, S.3.2 Classification systems, S.10 Accessibility and Clarity sub-concepts) is graphically highlighted.

4.3 Subprocesses of the statistical process with intermediate inputs and outputs

Information on general inputs and outputs for each edition of a statistical process is useful but to reach the goal of traceability that METAstat is persecuting, a major level of detail is needed. Each specific GSBPM subprocess should be described, including intermediate inputs and outputs and the methods applied. The structure that has been defined for the subprocesses in METAstat is inspired to the model proposed in Linking GSBPM and GSIM (UNECE, 2022). Indeed, for each subprocess, GSIM different types of inputs (e.g. *Core Input*, *Process support input*) and outputs (e.g. *Core Output*, *Process Metrics*) can be specified, while the *Process method* will provide the information on how the subprocess is carried out. The model proposed by UNECE will be supplemented by the Quality control actions and the Quality indicators (as

Process metrics) from SIDI-SIQual. They represent the tasks underlying the overarching Quality management process of GSBPM. Obviously, this structure would also allow to fill in many concepts and subconcepts of SIMS. A couple of examples are here after described:

- the intermediate *Core Output* produced by the GSBPM subprocesses 2.1 Design Outputs will be the definition of *Conceptual variables*. Such information can be used to fill in SIMS S.3.4 Statistical concepts and definitions.
- considering the GSBPM subprocess 4.3 Run collection (see figure 6 of Appendix 2), the *Process method* will provide the description of the data collection technique (corresponding to SIMS S.18.3 Data collection), while the quality control actions will report the activities carried out to reduce nonresponse and measurement error during data collection that correspond to what is asked in SIMS S.13.3 Non sampling errors. Overcoverage rate and unit nonresponse rate will be among the quality indicators produced as output (*Process Metrics*) of the subprocess and their values will fill in related fields of SIMS (13.3.1.1 Overcoverage rate, 13.3.3.1 Unit nonresponse rate).

Developing METAstat according to this conceptual model for all the subprocesses of GSBPM will allow to fill in automatically SIMS.

5. Concluding remarks

Several statistical authorities around the world are using international standards as GSBPM to document their production processes, usually as a first step towards modernisation and improvement of efficiency. Similarly, in many Countries, not only in the ESS, quality reports on statistical output are compiled according to SIMS and are used to provide information to users on quality of statistics disseminated or as a basis for quality assessment. There are also some experiences in which the documentation according to SIMS and GSBPM is managed in the same system but rarely one is obtained, even partly, from the other. To realise that automation, it is important to rely upon GSIM for the development of the conceptual model. In addition, it should be mentioned that a great contribution is due to the SIDI-SIQual approach that associates the quality actions and quality indicator explicitly to each GSBPM subprocess. This will allow to obtain a documentation of the statistical process that can be useful also as a base for the quality assessment. Finally, it should be noted that this complex model is possible also thank to the integrated management of reference, structural and terminology metadata.

References

Brancato Giovanna, Carbini Riccardo, Pellegrini Concetta, Signore Marina and Simeoni Giorgia (2006) Assessing quality through the collection and analysis of standard quality indicators: the Istat experience Proceedings of Q2006 European Conference on quality in survey statistics Cardiff, UK, 24-26 April 2006

Eurostat (2021) European Statistical System (ESS) Handbook for Quality and Metadata Reports — re-edition 2021 https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-21-021

Giorgia Simeoni (2013) Implementing ESS standards for reference metadata and quality reporting at Istat. UNECE Work Session on Statistical Metadata, 6-8 May 2013, Geneve, Switzerland https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP8.pdf

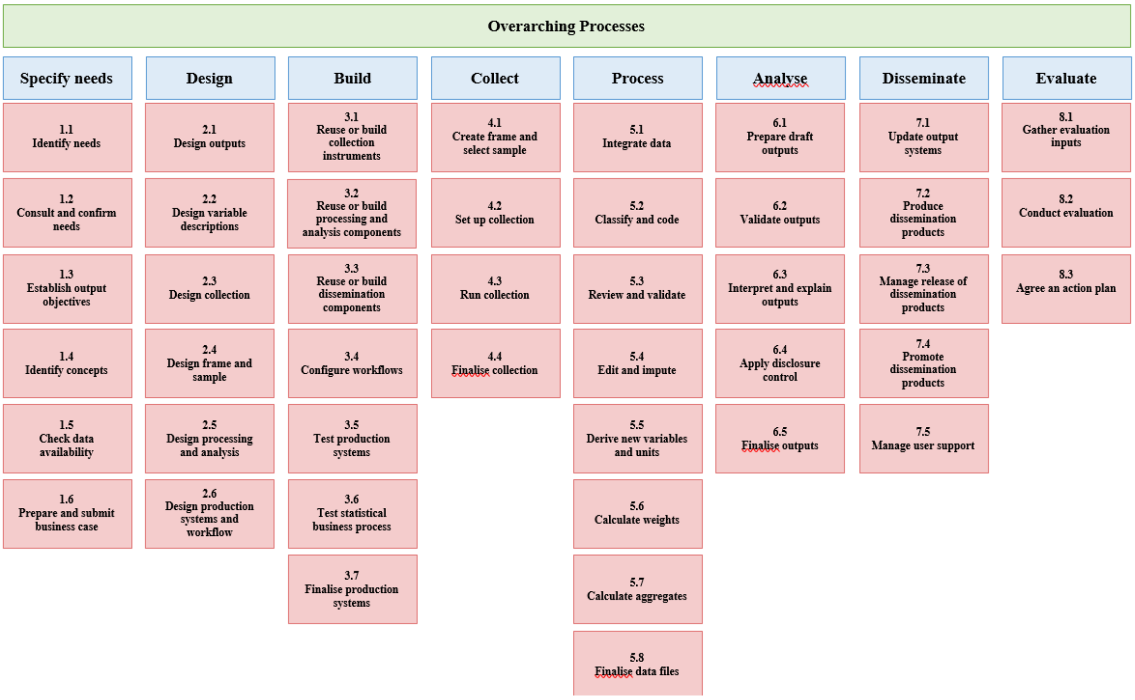
UNECE (2023). Generic Statistical Information Model GSIM 2.0 https://unece.org/statistics/modernstats/gsim

UNECE (2022). Linking GSBPM and GSIM (Version 1.0, January 2022) https://statswiki.unece.org/display/GSBPM/Information+flow+within+GSBPM+using+GSIM

UNECE (2019). Generic Statistical Business Process Model GSBPM (Version 5.1, January 2019) https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

Appendix 1. GSBPM and SIMS

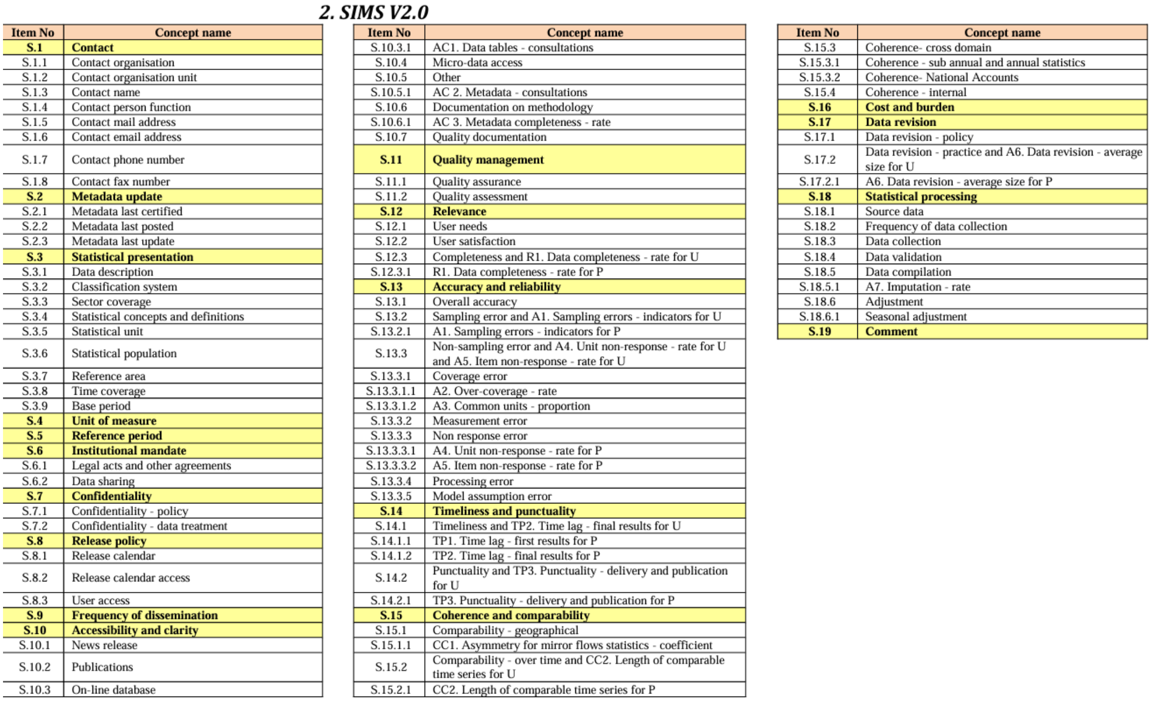
Figure 2. GSBPM 5.1



Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Figure 3. SIMS 2.0



2. SIMS V2.0

Item No	Concept name
S.1	**Contact**
S.1.1	Contact organisation
S.1.2	Contact organisation unit
S.1.3	Contact name
S.1.4	Contact person function
S.1.5	Contact mail address
S.1.6	Contact email address
S.1.7	Contact phone number
S.1.8	Contact fax number
S.2	**Metadata update**
S.2.1	Metadata last certified
S.2.2	Metadata last posted
S.2.3	Metadata last update
S.3	**Statistical presentation**
S.3.1	Data description
S.3.2	Classification system
S.3.3	Sector coverage
S.3.4	Statistical concepts and definitions
S.3.5	Statistical unit
S.3.6	Statistical population
S.3.7	Reference area
S.3.8	Time coverage
S.3.9	Base period
S.4	**Unit of measure**
S.5	**Reference period**
S.6	**Institutional mandate**
S.6.1	Legal acts and other agreements
S.6.2	Data sharing
S.7	**Confidentiality**
S.7.1	Confidentiality - policy
S.7.2	Confidentiality - data treatment
S.8	**Release policy**
S.8.1	Release calendar
S.8.2	Release calendar access
S.8.3	User access
S.9	**Frequency of dissemination**
S.10	**Accessibility and clarity**
S.10.1	News release
S.10.2	Publications
S.10.3	On-line database
S.10.3.1	AC1. Data tables - consultations
S.10.4	Micro-data access
S.10.5	Other
S.10.5.1	AC 2. Metadata - consultations
S.10.6	Documentation on methodology
S.10.6.1	AC 3. Metadata completeness - rate
S.10.7	Quality documentation
S.11	**Quality management**
S.11.1	Quality assurance
S.11.2	Quality assessment
S.12	**Relevance**
S.12.1	User needs
S.12.2	User satisfaction
S.12.3	Completeness and R1. Data completeness - rate for U
S.12.3.1	R1. Data completeness - rate for P
S.13	**Accuracy and reliability**
S.13.1	Overall accuracy
S.13.2	Sampling error and A1. Sampling errors - indicators for U
S.13.2.1	A1. Sampling errors - indicators for P
S.13.3	Non-sampling error and A4. Unit non-response - rate for U and A5. Item non-response - rate for U
S.13.3.1	Coverage error
S.13.3.1.1	A2. Over-coverage - rate
S.13.3.1.2	A3. Common units - proportion
S.13.3.2	Measurement error
S.13.3.3	Non response error
S.13.3.3.1	A4. Unit non-response - rate for P
S.13.3.3.2	A5. Item non-response - rate for P
S.13.3.4	Processing error
S.13.3.5	Model assumption error
S.14	**Timeliness and punctuality**
S.14.1	Timeliness and TP2. Time lag - final results for U
S.14.1.1	TP1. Time lag - first results for P
S.14.1.2	TP2. Time lag - final results for P
S.14.2	Punctuality and TP3. Punctuality - delivery and publication for U
S.14.2.1	TP3. Punctuality - delivery and publication for P
S.15	**Coherence and comparability**
S.15.1	Comparability - geographical
S.15.1.1	CC1. Asymmetry for mirror flows statistics - coefficient
S.15.2	Comparability - over time and CC2. Length of comparable time series for U
S.15.2.1	CC2. Length of comparable time series for P
S.15.3	Coherence- cross domain
S.15.3.1	Coherence - sub annual and annual statistics
S.15.3.2	Coherence- National Accounts
S.15.4	Coherence - internal
S.16	**Cost and burden**
S.17	**Data revision**
S.17.1	Data revision - policy
S.17.2	Data revision - practice and A6. Data revision - average size for U
S.17.2.1	A6. Data revision - average size for P
S.18	**Statistical processing**
S.18.1	Source data
S.18.2	Frequency of data collection
S.18.3	Data collection
S.18.4	Data validation
S.18.5	Data compilation
S.18.5.1	A7. Imputation - rate
S.18.6	Adjustment
S.18.6.1	Seasonal adjustment
S.19	**Comment**

Appendix 2 Conceptual models in METAstat

Figure 4. Example on general information of the statistical process

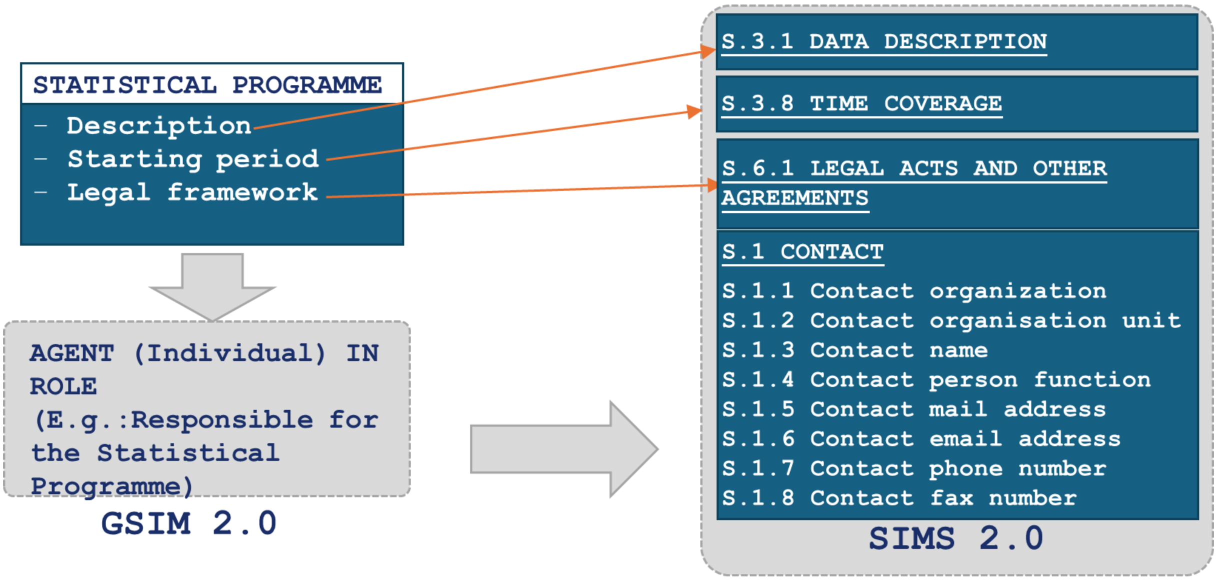

STATISTICAL PROGRAMME
– Description
– Starting period
– Legal framework
AGENT (Individual) IN ROLE
(E.g.:Responsible for the Statistical Programme)
GSIM 2.0
S.3.1 DATA DESCRIPTION
S.3.8 TIME COVERAGE
S.6.1 LEGAL ACTS AND OTHER AGREEMENTS
S.1 CONTACT
S.1.1 Contact organization
S.1.2 Contact organisation unit
S.1.3 Contact name
S.1.4 Contact person function
S.1.5 Contact mail address
S.1.6 Contact email address
S.1.7 Contact phone number
S.1.8 Contact fax number
SIMS 2.0

Figure 5. Example on inputs and outputs of the statistical process

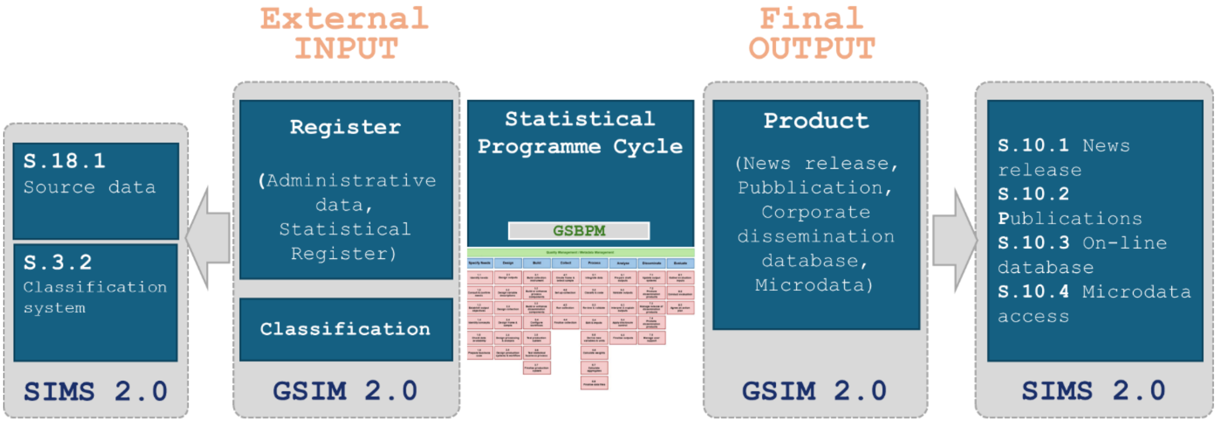
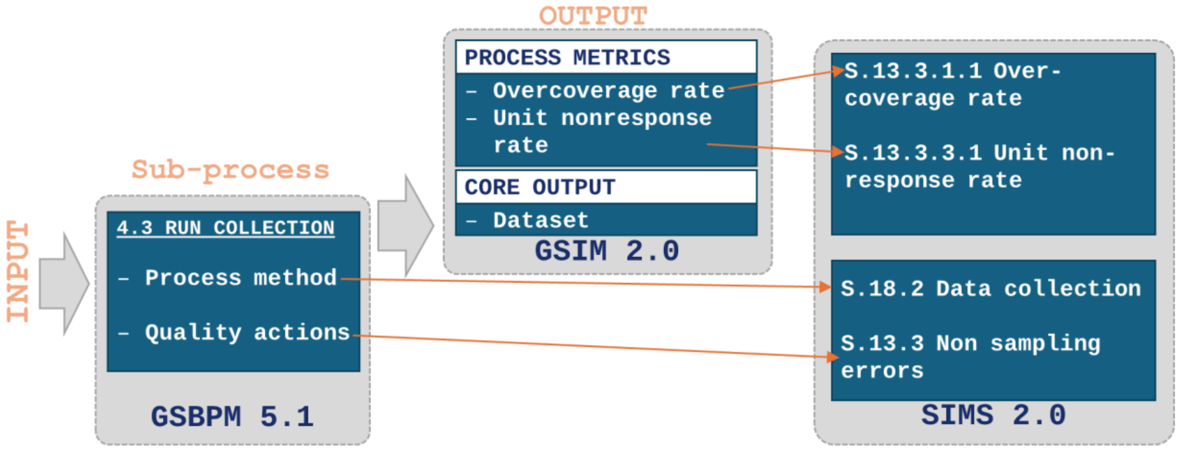


Figure 6. Example on a subprocess of the statistical process


OUTPUT
PROCESS METRICS
– Overcoverage rate
– Unit nonresponse rate
CORE OUTPUT
– Dataset
GSIM 2.0
S.13.3.1.1 Over-coverage rate
S.13.3.3.1 Unit non-response rate
S.18.2 Data collection
S.13.3 Non sampling errors
SIMS 2.0
Sub-process
INPUT
4.3 RUN COLLECTION
– Process method
– Quality actions
GSBPM 5.1