ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Work package K

Methodology and quality

Deliverable K4: Quality report template draft

Final version, 29.11.2019

Prepared by:
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
Sónia Quaresma (INE, PT)
Gabriele Ascari, Giovanna Brancato, Loredana Di Consiglio, Paolo Righi and Tiziana Tuoto (ISTAT, IT)
Piet Daas (CBS, NL)
Magdalena Six, Alexander Kowarik (STAT, AT)

Work package leader:

Alexander Kowarik (STAT, AT)
alexander.kowarik@statistik.gv.at
telephone : +43 1 71128 7513

Contents

Introduction

The structure of this template is taken from the widely known SIMS. The definitions and guidelines are based on the recently updated version of the EHQMR.

The members of the WPK went through each subconcept of the EHQMR, and for each subconcept we asked

- if a it is relevant for new data sources,
- if the definition of the (sub)concept has to be re-worded or
- if they can kept as they were.

If a subconcept was considered as not relevant, we deleted it. Since we kept the numbering of the subconcepts as in the EHQMR, the following subconcepts are not numbered consecutively.

When we had the impression that the existing subconcepts did not cover all relevant quality aspects for new data sources, we introduced new subconcepts. These new subconcepts are indicated by an "A" for "additional" in the subconcept number.

We came across the problem, that the SIMS is generally output-oriented, this means it measures the quality of Official Statistics. When it comes to new data sources, the output so far is almost never an Official Statistics, sometimes it is not even publishable.

In the case of the WPs of the ESSnet Big Data II, the “output” has often more the form of a throughput data set, which could further be used and processed. To avoid a problem with wording, we use the term "statistical output", which stands for the output of the WP, but does have to be a publishable statistical product.

You can immediately see the changes made by us since they are written in color.

This quality template is being tested by the WP members of the pilots track. It is used as basis for a questionnaire about quality issues for the pilots track intermediate meeting from 11th to 12th of December in Vienna. This is the reason that in some guidelines, you will find specific instructions what to report for the pilots track meeting. The feedback from the other WPs will be used to refine this quality report template.

Helpful Documents

Here you can find the updated but not yet officially published version of **the ESS handbook for quality and metadata reports (EHQMR)**

Here is the structure of the original SIMS
https://webgate.ec.europa.eu/fpfis/wikis/download/attachments/349740130/image2019-5-15_11-42-20.png?version=1&modificationDate=1557913341435&api=v2

S01 Contact

SIMS	Concept Name	Defintion	Guidelines
S.01	Contact	Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.01.1	Contact organisation	The name of the organisation of the contact points for the data or metadata.	Provide the full name (not just code name). of organisation responsible for the process and outputs (data and metadata) that are the subject of the report.
S.01.6	Contact email address	E-mail address of the contact points for the data or metadata.	Provide the email address(es) of the person(s) indicated as contacts. The address can be an individual e-mail address or a mailbox for the organisation to which the person has access.
S.01.7	Contact phone number	The telephone number of the contact points for the data or metadata.	Provide the telephone number(s) of the person(s) indicated as contacts.
Additional for big data			

S02 Metadata Update

SIMS	Concept Name	Definition	Guidelines
S.02	Metadata update	The date on which the metadata element was inserted or modified in the database.	(Information relating to this concept is provided by reporting on its sub-concepts.)

S.03 Statistical Presentation

SI MS	Concept Name	Defintion	Guidelines
S.03	Statistical presentation	Description of the statistical output.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.03.1	Data description	Main characteristics of the data set, referring to the statistical output.	Describe briefly the main characteristics of the data in an easily and quickly understandable manner, referring to the main variables. More detailed descriptions of the variables and how they were derived in S.03.4.
S.03.4	Statistical concepts and definitions	Statistical characteristics of statistical observations.	Define and describe briefly the main statistical variables that have been observed or derived. Indicate their types. Note that any difference between these variables and the variables desired by users is a relevance issue and is discussed in S.12. ***For the pilots track meeting:*** Indicate discrepancies, if any, from variables which were previously collected in a different way (e.g. via surveys).
S.3.5	Statistical unit	Entity for which information is sought and for which statistics are ultimately compiled.	Define the type of statistical unit about which data are available, e.g. enterprise, local unit, private household, person. If there is more than one type of unit, define each type. ***For the pilots track meeting:*** Summarize, if possible, the differences to units in traditional ways to collect data.
S.3.6	Statistical population	The total membership or population or "universe" of a defined class of people, objects or events.	Define the target population of the statistical units for which information is sought. Note that a difference between the target population and the population desired by users is a relevance issue and is discussed in S.12; and the difference between target population and the actual (frame) population is a coverage issue and is discussed in S13.3 If there is more than one type of population, define each type. ***For the pilots track meeting:*** Describe if there are any differences to the populations in the traditional surveys
S.3.7	Reference area	The country or geographic area to which the measured statistical phenomenon relates.	Describe the country, the regions, the districts, or the other geographical aggregates, to which the data refer. Identify any specific exclusions in the statistical data.

S.3.8	Time coverage	The length of time for which data are available.	State the time period(s) covered by the data, e.g. first quarter 2018, or quarters 2015-2018, or year 2018, or years 1985-2018. Note that any issues concerning comparability over time are discussed in S.15.
Additional for big data			

S.04 Unit of Measure

SIMS	Concept Name	Definition	Guidelines
S.04	Unit of measure	The unit in which the variables of the statistical output are measured.	The statistical data usually involves several units of measure depending upon the variables. Examples: • Country in which a SIM card is located at a certain time, • position of a ship at a certain time, • consumption of electricity in Watt/Kilowatt in a certain time span, • Classifying the negative or positive sentiments of a text input on a -1/1 scale
Additional for big data			

S.05 Reference Period

SIMS	Concept Name	Defintion	Guidelines
S.05	Reference period	The period of time or point in time to which the measured observation is intended to refer.	The value of a variable refers to a specific time period (for example, the last week of a month, a month, a fiscal year, a calendar year, or several calendar years), or to a point in time (for example, a specific day, or the last day of a month). The variables in a dataset may refer to more than one reference period. All reference periods should be stated Note that the difference, if any, between the target reference period(s) and the actual reference period(s) is an accuracy issue and should be discussed in S.13.3. Note that if frame population does not include all the units in the target population for the specified reference period, this is a coverage issue and should be discussed in S.13.3.
Additional for big data			

S.06 Institutional Mandate

SIMS	Concept Name	Definition	Guidelines
S.06	Institutional mandate	Set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.06.1	Legal acts and other agreements	Legal acts or other formal or informal agreements that assign responsibility as well as the authority to an agency for the collection, processing, and dissemination of statistics.	Describe the (legal) agreement and other forms of cooperation with the data owner which allows the NSI access to the data source. Describe which forms of reciprocity (not necessarily financial) does the NSI offer to the data source?
S.06.A	Data access and data transmission	Arrangements or procedures for data access and data transmission	Describe the arrangements, procedures or agreements for data access and data transmission. In particular, describe • Modes of data access (full access to raw data, access to pre-processed data, on-premise, off-premise) • In case of access to pre-processed data: transparency about the technological processes applied to the pre-processed data • Time and method of transmission • Time horizon of the cooperation - Is a long term access to the data guaranteed?

S.07 Confidentiality

SIMS	Concept Name	Definition	Guidelines
S.07	Confidentiality	A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.07.1	Confidentiality – policy	Legislative measures or other formal procedures which prevent unauthorised disclosure of data that identify a person or economic entity either directly or indirectly.	Describe all European or national legislation, or other formal requirements, that relate to confidentiality. Describe relevant policy (if any). Note that the existence of legislation and/or policy provides some assurance that methods necessary to assure confidentiality have been applied to the data. *European level* Summarise the commonalties and differences in national approaches to confidentiality policy..
S.07.2	Confidentiality - data treatment	Rules applied for treating the datasets to ensure statistical confidentiality and prevent unauthorised disclosure.	For aggregate outputs • Provide the rules that define a *confidential cell*. • Describe the procedures for detecting confidential cells, including checking for residual disclosure. • Describe the procedures for eliminating confidential cells, for example by controlled rounding, cell suppression, or cell aggregation. For micro-level outputs: • Describe the procedures that are used in protecting confidentiality.
S.07.A Additional for big data	Privacy	How privacy sensitive is the information coming from external data holders?	State which treatments are prescribed to satisfy privacy concerns

S.08 Release Policy

SIMS	Concept Name	Defintion	Guidelines
S.08	Release policy	Rules for disseminating statistical data to all interested parties.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.08.A Additional for big data	Release policy for Experimental Statistics	Rules for dissemination of experimental data or experimental statistical products.	State if there exists a publicly available policy for the dissemination of experimental statistics and if there exists a designated area at your NSI's homepage.

S.09 Frequency of Dissemination

SIMS	Concept Name	Defintion	Guidelines
S.09	Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period.	State the frequency with which the data are disseminated, e.g. monthly, quarterly, yearly. The frequency can also be expressed by using a code from the harmonised ESS code list so long as this is considered to be easily understandable by users.
Additional for big data			

S.10 Accessibility and Clarity

SIMS	Concept Name	Defintion	Guidelines
S.10	Accessibility and clarity	The conditions and modalities by which users can access, use and interpret data.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.10.6	Documentation on methodology	Descriptive text and references to methodological documents available.	List national reference metadata files, methodological papers, summary documents and handbooks relevant to the statistical process. For each item provide the title, publisher, year and link to on-line version (if any). ***For the pilots track meeting:*** List deliverables, reference metadata files, methodological papers, summary documents etc relevant to the process of deriving statistical data from raw data and - if already available - for producing statistical output using the statistical data.
S.10.7	Quality documentation	Documentation on procedures applied for quality management and quality assessment.	List relevant quality related documents, for example, other quality reports, studies. Cross reference to descriptions of quality procedures in other chapters, especially S.13. ***For the pilots track meeting:*** List also the deliverables, in which quality related issues are described.
Additional for big data			

S.11 Quality Management

SIMS	Concept Name	Definition	Guidelines
S.11	Quality management	Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.11.1	Quality assurance	All systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output.	Describe the quality assurance procedures specifically applied to the statistical process for which the report is being prepared, for example agreements with the big data providers, benchmarking, assessments, and use of best practices. Include descriptions of all forms of quality assessment procedures (self-assessment, peer review, compliance monitoring, audit) and when they most recently took place. Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail. Describe any ongoing or planned improvements in quality assurance procedures.
S.11.2	Quality assessment	Overall assessment of data quality, based on standard quality criteria.	Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail.
Additional for big data			

S.12 Relevance

SIMS	Concept Name	Defintion	Guidelines
S.12	Relevance	The degree to which statistical information meet current and potential needs of the users.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.12.1	User needs	Description of users and their respective needs with respect to the statistical data.	Provide: • a classification of users, also indicating their relative importance; • an indication of the uses for which users want the statistical outputs; • an assessment of the key outputs desired by different categories of users and any shortcomings in outputs for important users; • information on unmet user needs and any plans to satisfy them in the future; and • details regarding those quality components which do not meet user requirements.
S.12.3	Completeness	The extent to which all statistics that are needed are available.	Provide qualitative information on the extent to which content requirements in relevant legislation, regulations and guidelines are met. Provide information on the extent to which user needs related to content are satisfied. Provide values of indicator R1 Data completeness rate, for each required data item for each relevant regulation/ guideline at producer/user level of detail as appropriate. In the case where the indicator refers to data sent to Eurostat, this indicator can be compiled by Eurostat. *European level* Summarise across countries the extent to which ESS requirements for data items are met
S.12.A Additional for big data	Added Value through new data source	The potential added value of a new data source to an existing statistical product.	Describe if and how the usage of a new data source provides an added value to an already existing statistical product. E.g., this could be more detailed data on particular subgroups, or information on grid level instead of district level or the potential replacement of questions of a survey through information of the new data source.

S.13 Accuracy and Reliability

SIMS	Concept Name	Definition	Guidelines
S.13	Accuracy and reliability	Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.	(Information relating to accuracy is provided by reporting on S.13 sub-concepts. Information on Reliability is reported in S.17 Data Revision).
S.13.1	Overall accuracy	Assessment of accuracy, linked to a certain data set or domain, which is summarising the various components.	Describe the main sources of random and systematic errors in the statistical outputs and provide a summary assessment of all errors with special focus on the impact on key estimates. The bias assessment can be in quantitative or qualitative terms, or both, and may be expressed as bias risk. It should reflect the producer's best current understanding (sign and order of magnitude) and include actions taken to reduce bias. *European level* Provide a summary picture of accuracy across countries. The emphasis placed on various types of errors should depend upon the error profile of the respective process. For repetitive processes, describe how accuracy is developing over time and what efforts are underway to improve accuracy from an ESS perspective. *Comment for the track meeting:* There is a tendency to focus on the micro-level here. please include in this and subsequent sections that reporting at the group or aggregated level can/should be done when the units can not be identified. In general, one should be able to repport any quality issues when working with event-based Big Data sources!!!
Additional for big data			

S.13.2	Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.	State whether sampling error is relevant. If probability sampling is used: • for user reports, provide the range of variation of the A1 indicator among key variables at user report level of detail; • for producer reports, provide the range of variation of the A1 indicator among key variables at producer report level of detail; • indicate the impact of sampling error on the overall accuracy of the results; • state how the calculation of sampling error is affected by imputation for nonresponse, misclassifications and other sources of uncertainty, such as outlier treatment. If non-probability sampling is used, provide an assessment of representativity and risk of sampling bias. *ESS level* If probability sampling is used: • present sampling errors for key estimates across countries; • indicate which country to country differences are significant and which are not; • for a repetitive survey, describe at least broadly the trends in sampling error over time • provide sampling errors for ESS level estimates.
S.13.3	Non-sampling error	Error in estimates which cannot be attributed to sampling fluctuations	Summarise the most important aspects of coverage, measurement, non-response, processing and model assumption errors. Discuss the corresponding bias risks and actions undertaken to reduce them.
	A4. Unit non-response - rate (U)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a user report.	

	A5. Item non-response - rate (U)	The ratio of the in-scope (eligible) units that have not provided a particular item and the in-scope units that are expected to provide that particular item, at a level of detail appropriate for a user report.	
S.13.3.1	Coverage error (P)	Divergence between the population of the Big Data source and the target population.	Provide information on the frame and its sources and actions performed to gather the population impacting on coverage (e.g. webscraping). Provide an assessment, whenever possible quantitative, of overcoverage and undercoverage, including an evaluation of the bias risks associated with the latter. Describe actions taken for reduction of undercoverage and associated bias risks
S.13.3.1.1	A2. Overcoverage – rate (P)	The proportion of units accessible via the frame that do not belong to the target population.	Report A2, Overcoverage - rate
S.13.3.2	Measurement error	Measurement errors are errors that occur during data capture and cause recorded values of variables to be different from the true ones	The main sources of measurement error should be reported and assessed. Their description should be accompanied by any available analysis, otherwise by the producer's best knowledge. Where available and relevant describe: • identification and general assessment of the main sources of measurement error, including errors arising from data acquisition; • efforts made in questionnaire design and testing, information on interviewer training and other work on error prevention; • errors in measurement instruments (meters, satellites,...) • *results of assessments based on comparisons with external data, re-interviews or experiments;* • *results of indirect analysis, for example, of the editing phase; and* • actions taken to correct measurement errors.

S.13.3.3	Nonresponse error	Nonresponse errors occur when the Big Data source fails to collect one or all the variables for units belonging to the domain covered by the source	Provide qualitative/quantitative assessments of unit nonresponse and highlight the units that are most subject to nonresponse Highlight the variables that are most subject to item nonresponse Provide a qualitative/quantitative assessments of the bias associated with nonresponse, comparing response rate for different sub-groups or distribution of auxiliary variables known for respondents and non-respondets (etc.) Provide a breakdown of nonrespondents according to cause for nonresponse mainly focusing on unit dependent cause and data collection tools cause. Define a stategy for reducing nonresponse during data collection and follow-up. Implement an estimator adjusted for nonresponse . *European level* ***(not needed for the pilots track meeting)*** Provide a qualitative/quantitative assessments of unit and item nonresponse across countries.
S.13.3.3.1	A4. Unit nonresponse - rate (P)	The ratio of the number of units with no information or not usable information to the total number of in-scope (eligible) units, at a level of detail appropriate for a producer report.	Report A4: Unit nonresponse rate overall and at a level of detail appropriate for a producer report.

S.13.3.4	Processing error	The error in final data collection process results arising from the faulty implementation of correctly planned implementation methods, e.g., algorithms used to transform the data or extract information from raw data.	If processing errors are significant, identify the main issues regarding them. Present an analysis of processing errors, where available, otherwise a qualitative assessment. Report their extent, and impact on the outputs, of the most significant types of error. Include descriptions of linking and coding errors, if applicable. Where mistakes relating to programming or publishing have occurred, corrective measures taken as well as actions for avoiding them in the future should be reported. Example: For web data sources: Setting up a pipeline assures processing is comparable over time. Because texts were processed, the final results were highly affected by the various choices of text processing made.
S.13.3.5	Model assumption error	Error due to models used in the statistical production.	Describe process specific models, for example, as needed to define the target of estimation itself and models used for transformation of data into statistical data. Provide an assessment of the validity of each model. Descriptions of models used in treatment of specific sources of error should be presented in the section dealing with those errors. The assessment of the models used in treatment of specific sources of error should be presented in this section. Discuss the trade off between the need to use proper model that can change over time (accuracy) and the use a constant model in order to ensure comparability over time
13.3.5.A Additional			

S.14 Timeliness and Punctuality

SIMS	Concept Name	Defintion	Guidelines
S.14	Timeliness and punctuality	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.14.1	Timeliness	Length of time between data availability and the event or phenomenon the data describe.	Outline the reasons for the time lag. Outline efforts to reduce time lag in future. ***For the pilots track meeting*** Describe the envisioned time lag for producing statistical output from/with the help of a new data source. Describe if the use of the new data source has the potential to decrease the time lag as it exists at the moment for already existing statistical products.

S.15 Coherence and Comparability

SIMS	Concept Name	Defintion	Guidelines
S.15	Coherence and Comparability	Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.15.1	Comparability – geographical	The extent to which statistics are comparable between geographical areas.	Describe any problems of comparability between regions of the country. The reasons for the problems should be described and as well an assessment (preferably quantitative) of the possible effect on the output values. Give information on discrepancies from the ESS/ international concepts, definitions, with reference to other chapters for more details.
S.15.2	Comparability – over time	The extent to which statistics are comparable or reconcilable over time.	Provide information on possible limitations in the use of data for comparisons over time. Distinguish three broad possibilities: 1. There have been no changes, in which case this should be reported. 2. There have been some changes but not enough to warrant the designation of a break in series. 3. There have been sufficient changes to warrant the designation of a break in series. For the pilots track meeting: Additionally, provide information about the comparability over time of the technological processes which produce the data, of the data access and changes in the covered population over time. Give also an assessment how the the comparison over time will develop in the future.
S.15.3	Coherence-cross domain	The extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.	An analysis of incoherence should be provided, where this is an issue of importance.

S.15.4	Coherence – internal	The extent to which statistics are consistent within a given data set.	Each set of outputs should be internally consistent. If statistical outputs within the data set in question are not consistent, any resulting lack of coherence in the output of the statistical process itself should be stated as well as a brief explanation of the reasons for publishing such results.
S.15.A.1	Coherence - with existing information/ Official Statistics	The extent to which information / statistical output from new data sources is consistent with information /statistical output from traditional data sources.	Provide information if it is meaningful to compare the information gained from new data sources with information from traditional data sources and if so, how consistent the information /statistical output gained from new data sources is with the one from traditional data sources.
S.15.A.2	Comparability - between information from several distinct new data sources	The extent to which information from several distinct new data sources is comparable among one another.	If you have raw data from several distinct new data sources, provide information how comparable the respective raw data sets and the information derived from them are among one other. Examples: MNO data from several mobile operators, smart meter data from several electricity providers

S.16 Cost and Burden

SIMS	Concept Name	Defintion	Guidelines
.16	Cost and burden	Cost associated with the collection and production of a statistical product and burden on respondents.	***Cost*** Provide annual operational costs of the process, with breakdown by major cost component. Describe recent efforts to improve efficiency and comment on the extent to which information and communication technology is used. *European level* Describe recent initiatives and efforts to improve efficiency at the European level. ***Burden*** Provide an estimate of the respondent burden imposed by the process. Describe all the means taken to minimise burden. *European level* Describe recent initiatives and efforts to minimise burden at the European level.
S.16.A Additional for big data	Potential savings in cost and burden	Description how the new data source might influence cost and burden in the future	Provide an overview how the new data source could be deployed in the future to save the NSIs cost and/or decrease the respondent burden. Provide a qualitative description of the additional efforts for the NSI and the data owners.

S.17 Data Revision

SIMS	Concept Name	Definition	Guidelines
S.17.1	Data revision – policy	Policy aimed at ensuring the transparency of disseminated data, whereby preliminary data are compiled that are later revised.	Describe the data revision policy applicable to data output from the statistical process being reported. In so far as they are relevant to the process being reported, summarise the general procedures for treatment of planned revisions, benchmark revisions, unplanned revisions, and revisions due to conceptual and/or methodological changes. *European level* Describe the data revision policy and procedures at European level.

S.18 Statistical Processing

SIMS	Concept Name	Definition	Guidelines
S.18	Statistical processing	(Defined by its sub-concepts)	(Information relating to this concept is provided by reporting on its sub-concepts.)
S.18.1	Source data	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	Indicate if the data are based on a survey, an administrative data source, multiple data sources, big data source (machine generated, human sourced, process mediated), e.g., web data, and/or macro-aggregates. Refer to the accreditation document of the data source, if applies. In the event of multiple data sources or macro-aggregates, reference each source and indicate how they are combined. For each survey source, summarise the sample design, cross referencing the descriptions of the target and survey populations, presented in S.03.6. For each ~~administrative~~ data source, summarise the source, its primary purpose, and the most important data items acquired. Information in which form the metadata for the new data source is available, where it can be found, and if it is updated on a regular basis. *European level* Provide an overview of the sources used across countries.
S.18.2	Frequency of data collection	Frequency with which the source data are collected.	Indicate the frequency of data collection (e.g. monthly, quarterly, annually, or continuous).

S.18.3	Data collection	Systematic process of gathering data for official statistics.	*For each survey data source:* • describe the method(s) used to gather data from respondents; • annex or hyperlink the questionnaires(s). *For each administrative data source* • describe the acquisition process and how it was tested. *For all sources* • describe the types of checks applied at the time of data entry. *For big data sources* describe the methods used to collect the data; add hyperlink if it is web data or name of the API used to collect the data. *European level* Provide a summary of the commonalities and differences in the collection methods, questionnaires and checks used in different countries.
S.18.4	Data validation	Process of monitoring the results of data compilation and ensuring the quality of statistical results.	Describe the procedures for checking and validating the source data and how the results are monitored and used. Describe the procedures for validating the aggregate output data (statistics) after compilation, including checking coverage and response rates, and comparing with data for previous cycles and with expectations. List other output datasets to which the data relate and outline the procedures for identifying inconsistencies between the output data and these other datasets. Define the linkage method for big data sources and other data sources used for validation. *European level* Provide a summary of the commonalities and differences in the validation methods used by countries.

S.18.5	Data compilation	Operations performed on data to derive new information according to a given set of rules.	If there is missing data, give detailed description of the methods used for imputation. For big data sources, e.g., web data, indicate the reason why data were not collected (technical issues etc.). Describe the procedures for imputation, the most common reasons for imputation and imputation rates within each of the main strata. Describe the likely impact of imputation. Describe the procedures for adjustment for non-response and the corrections to the design weights to account for differences in response rates. Describe the calculation of design weights, including calibration (if used). Describe the procedures for combining input data from different sources.
S.18.5.1	A7. Imputation – rate	The ratio of the number of replaced values to the total number of values for a given variable.	Provide values of indicator A7 Imputation – rate

S.19 Comment

SIMS	Concept Name	Defintion	Guidelines
S.19	Comment	Supplementary descriptive text which can be attached to data or metadata.	Provide any information that is pertinent to the report but does not fit under any of the other concepts, or to repeat key issues, or to make reference to annexes that might be attached to the report.