ROMANIA

Reimbursable Advisory Services Agreement on
Romania Capacity Building for Statistics (P167217)

OUTPUT No. 6d

Report on advisory services provided to Recipient on the

Methodology for a continuous and real time evaluation of the quality of data collected from administrative and geo-spatial sources for the inter-census periods

May 2022

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Project co-financed from the European Social Fund through the Operational Programme for Administrative Capacity 2014-2020

Disclaimer

This report is a product of the International Bank for Reconstruction and Development / the World Bank. The findings, interpretation, and conclusions expressed in this paper do not necessarily reflect the views of the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work.
This report does not necessarily represent the position of the European Union or the Romanian Government.

This report has been delivered in May 2022 under the Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217) signed between the Romanian National Institute of Statistics and the International Bank for Reconstruction and Development on September 17, 2019. It corresponds to Output 6d under the above-mentioned agreement.

Acknowledgements

This report was prepared by Tiziana Tuoto and Carlo Vaccari under the coordination of Michael Wild, Senior Statistician, World Bank with the support of local team expert Antoniade Ciprian Alexandru. The team would also like to express its gratitude to government officials of the National Institute of Statistics (INS), to Silvia Pisică (Project manager), Lavinia Bălteanu (Process responsible), Ruxandra Moldoveanu (Expert), Laura Ichim (Expert), Manuela Vlaicu (Expert), Mihaela Ram-Chand (Expert), Florica Frîncu (Expert) for their constructive collaboration.

Table of Contents

Table of Figures

List of Acronyms

INS	National Institute of Statistics
NSO	National Statistic Office
RAS	Reimbursable Advisory Services
WB	World Bank

Executive summary

The purpose of this report is to provide guidance and recommendations for a continuous and real time evaluation of the quality of data collected from administrative and geo-spatial sources in the inter-census periods with the aim of creating a complete registers and census - based frame for the calculation of population counts. The document provides both qualitative and quantitative measures for quality evaluation. Three checklists and corresponding templates, provided in the appendices, are the operating tools for quality evaluation. The document is intended as a guide for the usage and the compilation of the checklists.

This is part of the deliverables under the Reimbursable Advisory Services (RAS) Agreement on Romania Capacity Building for Statistics (project No. P167217). The project is implemented by the National Institute of Statistics with support from the World Bank.

This report is divided into four (4) sections and includes seven appendices which are an integral part of the report.

The background section introduces the concept of “Stage” and “Hyper-dimension” used in the report. Section A provides a description of the usefulness of administrative and geo-spatial data during inter-census periods, focusing on the Quality dimensions of the sources and on the tools proposed to assess the quality.

Section B provides a description of the metadata hyper-dimension and the checklist proposed to assess metadata quality. Section C presents recommendations for INS to identify and measure the different types of errors that can occurs managing a census/register-based frame.

The appendices complete the report with the checklist and templates proposed to INS to assess data and metadata quality. Three checklists were also compiled together on a pilot source by the INS-WB working group, to complete the hands-on technical assistance and verify together the correspondence of the tools to the needs of the INS.

Basic Glossary:

Frame: Any list, material or device that delimits, identifies, and allows access to the elements of the target population. A statistical register is a specific example, as well as an administrative register.

- *Population register*: A statistical register and a frame of persons usually resident (however defined) in a country. Additionally, it often provides some demographic characteristics of individuals

Unit: The smallest entity to which any data item refers. Units may refer to individual persons, households, buildings, or dwellings

- *Administrative unit*: The units for which administrative data are recorded. These may or may not be the same as those required for the statistical output (which are referred to as statistical units). In some of the literature (e.g., Zhang 2012), the term '*object'* is used to refer to the units within an administrative dataset. The term is used to distinguish between units in the administrative data and the statistical units after this data has been transformed in some way. This is particularly relevant in cases where the unit (or 'object') in the administrative register differs from the target statistical unit. For example, where a tax register, where the units of a yearly tax returns (i.e., the same person may make several returns in one or multiple years), is converted into individual 'people'
- *Composite unit*: Units composed by one or more individual units. Example: household, which is composed by persons (in business statistics we have also enterprise/company composed by one or more establishments).
- *Relationship between different target units*

Variable: A socio-demographic or economic characteristic/attributes relating to an administrative or statistical unit for which information is required for the purpose of the frame

- *Core variables*: variables that identify and allow access to the elements of the target population. Examples: for the update and maintenance of an admin/census-based frame, core variables are PIN, name, surname, gender, place, and date of birth, contact info like residence address (preferably geo-referenced), municipality, email, phone. In this way you may know the basic demographic statistics (age and gender composition of the population) as well as contact info for sample surveys. A specific discussion is needed for the variable "resident status" which specifies the characteristic of belonging to the target population.
- *Non-core variables*: any other variables that do not contribute to identify and access to the elements of the target population. Examples: for the update and maintenance of a register/census- based frame, non-core variables are educational status, civil status, employment status, etc. These variables can be obtained by administrative sources, but they are not essential to identify target units and contact them. Sometimes the available information in the administrative sources do not perfectly correspond to the statistics definitions/requirements, which need to be derived by sample surveys.
- *Derived variable*: A new variable formed by using the data from other variables. The variable "resident status" is often a derived variable in the statistical frame

Background:

A framework for assessing data quality with stages and hyper-dimensions

The quality of a frame produced using administrative sources is particularly difficult to assess and/or measure due to the complexity and multi-dimensionality of the data used. Some factors affecting quality are not quantitatively measurable. We can distinguish between assessing quality, meaning a qualitative evaluation, and measuring quality – meaning attaching a quantitative metric to this evaluation of quality. Where it is not possible to produce indicators for quantitative measurement, or where they have not yet been developed, we recommend a qualitative assessment of their impact on quality.

A frame that uses census and administrative sources relies partially on data that were produced outside of the statistical system, in a different organization over which the INS usually has no control. For this reason, the impact of using these outside *sources* must be considered carefully. For this reason, we can apply a quality framework prior to the statistical usage of the administrative data, i.e., at the input stage, to determine if an administrative data source can be used for statistical purpose and how. To clarify and systematize the stages (hyper-dimensions) of a framework for quality assessment of secondary data, we can refer to some international standard.

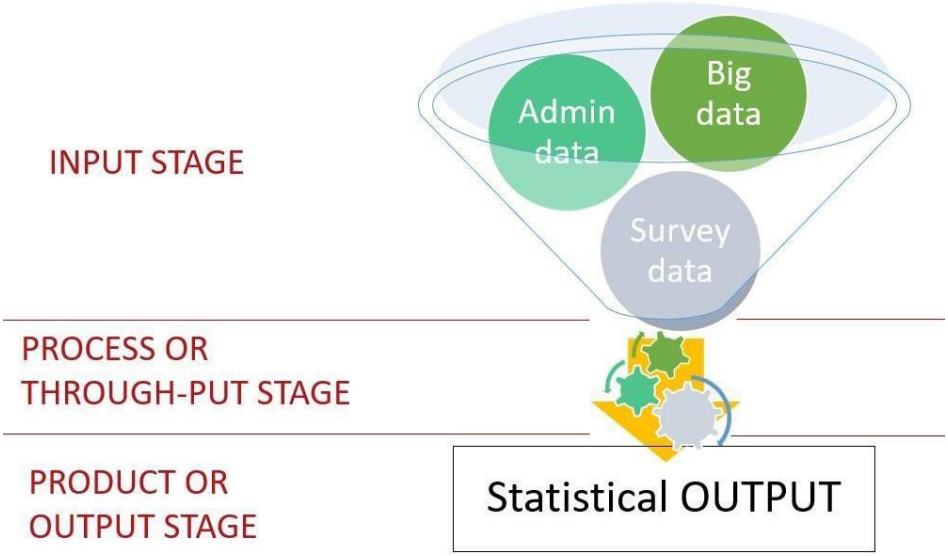
The framework for quality assessment has some desired characteristics. In particular, the assessment is likely to start with no data in hand. There should be an initial step in the assessment framework that guides the initial acquisition of the data and can be applied with or without data in hands. In addition, the assessment should offer a tool that takes the several statistical potential usages of administrative data into consideration. Finally, the assessment should offer commonalities with the dimensions found in recognized quality assessment frameworks.

Stages and hyper-dimensions of quality assessment

To ensure a complete and easy to follow quality assessment, we can consider broad stages of the administrative sources' lifecycle. They are applicable regardless of the intended use. While statistical process design is never entirely linear, thinking of how to carry out quality assessment in this way should enable us to quickly identify the key quality considerations most relevant to each circumstance. In addition, this allows us to customize each stage to cover several quality dimensions and associated quality indicators.

The main stages of the secondary data lifecycle are depicted in figure 1.

Figure 1 - The main stages of the data lifecycle



The Stages are:

(a) **Input Stage**: at this stage we recognize the *Source* hyper-dimension and the *Metadata* hyper-dimension,and we need to describe both source and metadata-based quality assessment of new or re-suppliedadministrative sources to be used. This Stage does not require INS to be in possession of the actual data, but it is crucial for the Stages that follow. In cases when the data are already available, the completion of the Source and Metadata quality assessment is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the administrative data sourcegiven the intended uses.

(b) **Process Stage**: at this stage we aim at assessing the quality of the raw administrative data supplied by administrative authorities (data suppliers) and we recognize the *Data* hyper-dimension and the *Process* hyper-dimension. This will require NSOs (INS in our case) to validate the data supplied against the lessons learned from the Source Stage. As well as basic validation, this Stage includes any processing required to establish the qualityof the data supplied vis-à-vis what was expected and comparisons with alternative sources. The Process hyper-dimension aims at assessing the quality of the several processes often carried out on administrative data sources, to transform the data for statistical usage and/or to improve quality. The processes identified include:

(i) Record linkage,
(ii) 'Signs-of-life' methodology,
(iii) Conflict resolution/decision between sources, and
(iv) Editing and imputation.

The transformations to which administrative data are subject to in order to guarantee their statistical usage are not the subject of this report. Some of them will be discussed in the Output 6b - *Methodology on sampling methods when switching to other type of sample frame (e.g., from master sample to register based frame) for surveys and statistics in the inter-census periods*, where quality aspects will be highlighted.

(c) **Output Stage**: The overall quality assessment of the statistical outputs produced using administrative data. This might not be conceptually that different from the assessment of the outputs of traditional surveys and thus, it does not make the subject of the present report.

A. Analysis of administrative and geo-spatial sources usefulness in relation to the proposed purpose

1. Identify the intended use of the administrative and geo-spatial sources

In inter-census periods, we can identify three main purposes:

a) update the register/census-based frame with respect to the coverage of the target units and their basic information (core variables).

b) update the register/census-based frame with respect to the other useful information available inthe sources for producing required statistics (non-core variables).

c) derive counts related to the target units broken down by some basic core variables.

The National Statistical Institute can concentrate on the first purpose, since it is the most important one andadministrative sources really play a crucial role for it; on the other side we will provide insight on the secondpurpose where it is a by-product of the first one, taking in mind that statistics for the non-core variables cannot always be derived by register/census-based frame while sometimes they need dedicated sample surveys (this is a common case for the employment statistics, for instance).

The third purpose, i.e. derive counts for some basic variables, is a by-product of the first one, i.e. having an updated register/census-based frame to be used for different purposes. However, the derivation of aggregated counts can be considered also a preliminary stage compared to the full availability of an updated register/census-based frame and some counts can be achieved even at an early stage of the register/census-based frame under some simplified working assumptions.

2. Quality dimensions for the source

According to the most accepted and used international standards, the quality dimensions for data sources can be summarized in Table 1. It is useful to investigate the quality dimensions in Table 1 even before the INS has direct access to the data sources, in a discovery phase of the data acquisition stage.

Figure 2 - Quality dimension of Source stage

	QUALITY DIMENSION	DEFINITION
SOURCE STAGE	Relevance and Accuracy	The degree to which the administrative data source meets the needs of the register/census-based frame. This covers the overlap between the target population, concepts, and definitions (relevance) and the degree to which the data correctly describe the phenomena they weredesigned to measure (accuracy).
	Timeliness	The lapse between the end of the reference period to which the information pertains and the date on which the information becomes available to the NSO.
	Coherence and Comparability	The degree to which the administrative source can be successfully combined with other sources, including linkability.
	Accessibility and Interpretability	The ease in which the NSO can obtain the administrative data, covering the impact of any restrictions, privacy and security, public acceptability of the use, the ease of data transfer and receipt, and the availability of metadata.
	The Institutional Environment	Organizational factors affecting the data supplier's capacity to supply data to the quality expected. Covering the strength of the relationship, previous experience, existence offormal agreements, risks associated with the status of the supplier and the supplier's quality standards.

3. Identify/design a proper tool for assessing the usefulness of the sources

The analysis of the usefulness in relation to the purpose of updating and maintaining the register/census- based frame can be assessed even without the data source in your hands, and this is a great advantage in order to focalize and priorities with respect to the data acquisition phase.

A useful tool for assessing the usefulness of a data source is provided in Appendix A. "**Quality checklist for the Input phase – Source**". This tool is also useful when the purpose of the usage is still not clear at the beginning and when a National Statistical Office is still investigating/inquiring through a data acquisition stage. This tool can be adapted to the INS specific case.

In Appendix A1 we provide the **Quality checklist for the Input phase – Source** filled in for the Source "Tax Agency" by the WB and INS team during a dedicated virtual meeting.

We recommend performing the evaluation of the data sources in three subsequent steps using:

1) the Source checklist;

2) the Metadata checklist; and finally,

3) the Data checklist (a more in-depth quality assessment).

In general, a positive outcome in one step would indicate to pursue the next step. The Source and Metadata checklists can be performed even without accessing the actual data. The check lists are a working tool for translating in main figures the quality concepts. In the checklists some concepts can be omitted (e.g. the Coherence and Comparability are not included in the Source checklist), other can be stressed, and the checklist can be adjusted according to specific needs and situations.

As anticipated, each step covers several quality dimensions and associated quality indicators. In the checklists, a score must be assigned to each quality indicator with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a rank chosen between high, medium, and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation.

As already mentioned, the checklist provides a useful tool even when the INS is exploring the data with no specific intended use in mind and is doing so to discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided.

It is worthwhile noting that the summary table at the end of the checklist provides precious assistance to identify weaknesses and strengths of the considered sources.

B. Maintenance of metadata associated with administrative data

1. The metadata hyper-dimension of the administrative and geo-spatial sources

At the Input stage, we can have a metadata-based quality assessment of new or re-supplied administrative sources to be used, even when INS is not in possession of the actual data. As anticipated, when the data arealready available, the completion of the Metadata quality assessment is also recommended as it contributesto document quality aspects.

The quality dimensions to consider for the Metadata hyper-dimension are related to the level of information provided to assess the Interpretability, the Relevance, and the Coherence/Accuracy of the Source.

A useful tool for quality assessment of the Metadata hyper-dimension is still a checklist. Appendix B providesa **Quality checklist for the input phase – Metadata**, which can be customized in order to take into account INS specificities.

In Appendix B1 we provide the **Quality checklist for the input phase – Metadata** filled in for the Source "TaxAgency" by the WB and INS team during a dedicated virtual meeting.

2. A template to assess the quality of Data

Once the data is in the possession of the INS, at the Process stage, a template might help in collecting and organizing the data (hyper-dimension) quality assessment.

Appendix C provides a template which is designed to capture the key aspects of quality for a given dataset in an organized way. It is a part of the larger assessment of data quality. This template needs to be customized on the specific INS needs, not every single box in this template could be useful. In addition, the level of detail might be tailored to the needs of the particular project or investigation. However, we recommend keeping in mind that information that the INS team enters into this template will be very valuable to any other people who use the dataset in the future. For this reason, as in the previous checklists, we recommend using easily understandable language, include all details and definitions, and do not assume too much of the other readers.

The template also contains elements for the identification and measurement of various types of errors. Again, the level of detail at which measuring and reporting the errors can be tailored to the specific INS situation.

The quality dimension to consider in the Data stage can be summarized in figure 3.

Figure 3 - Quality dimension at the Data Stage

	Quality Dimension	Definition
DATA STAGE	Validation and Harmonization	The data files provided to the NSO are in a readable format. Further data validation and harmonization arrangements are in place upon data transfer to the NSO. This is done toconfirm that the expected variables/units/reference period/formats have been supplied and to ensure data processing by the NSO is consistent across register / census-based frame use cases.
	Accuracy and Reliability	The accuracy, completeness (for variables andpopulation coverage) and coherence of the data supplied matches the requirements of the specific use case for which it will be used. Comparisons with alternative sources reveal acceptable levels measurement or representative errors.
	Timeliness and Punctuality	The timeliness and punctuality of the data supplied matches the requirements of the specific use-case for which it will beused.
	Linkability	Adequate linkage variables are available (i.e.,either common unique identifiers or a combination of variables which enable identification) and these are of sufficient quality to enable data linkage.

A shorter version of the template is the "**Quality checklist for the acquisition phase - Data**" provided in Appendix D. It collects the most relevant aspects of the previous templates, in particular those related to the main types of errors that might compromise the analysis based on the administrative and geo-spatial data. In Appendix D1 we provide the **Quality checklist for the acquisition phase – Data** filled in for the Source "Tax Agency" by the WB and INS team during a dedicated virtual meeting.

C. Identification and measurement of various types of errors

A census/registers-based frame must meet quality criteria in terms of updating, coverage and accuracy of theinformation contained therein. The ideal frame should meet the following requirements:

- To be made up only of the units belonging to the population of interest at the time of the survey.
- To include each unit of the population only once.
- To contain updated and correct data regarding the identification information (name and address)and any basic descriptive information (the core variables) of the units.

These characteristics allow the usage of the frame from a strictly theoretical sample perspective, as well as for the reduced target of counting target units by some specific breakdowns.

Possible situations of deviation from the ideal situation are due to "errors" that we might classify as:

- **under-coverage**, i.e., some units of the population are not included in the frame.
- **over-coverage**, i.e., some units of the frame are non-existent and / or do not belong to the target population; a special case of over-coverage is the *duplication* of some units, if some elements of thepopulation are present several times in the frame.
- **inaccurate and missing values** in the core variables.
- clusters of units, when some elements of the frame contain clusters of elements of the population, e.g. we have a frame of households whilst the target population are the individuals.

Checklists and templates presented in the previous sections contain relevant indicators to measure the aforementioned errors.

It is worthwhile noting that the under-coverage and over-coverage concepts are strictly related to the location of the units. Actually, we can distinguish cases completely missed by the frame (pure under-coverage) as well as cases mis-located, i.e., a person is assigned to an address in Sofia while he/she is actually resident in a different town/village. The latter cases represent local under-coverage compensates at nationallevel by corresponding over-coverage somewhere else. These local under-coverage and local over-coverageneed to be considered (and measured) carefully both for the main purpose of establishing and maintaining acomplete frame and for the reduced purpose of counting population size by some break-down. Hence, since coverage errors need to be reduced as much as possible, it is appropriate to include in the quality assessment all the geo-spatial information that can be derived by the secondary sources, for a clear evaluation of their impact in terms of coverage errors.

D. Recommendations

Several recommendations are presented to apply for administrative data sources:

a) Identify relevant and promising administrative sources for use in the census/register-based frame update.

b) Set out clearly the required target population, variables, and concepts, along with the anticipated outcomes for using an administrative source on which to base the assessment.

c) Understand the restrictions and challenges to acquiring and integrating administrative sources, includingwhere changes may be needed to the INS's methods, processes, and computing systems.

d) Build and maintain clear and comprehensive metadata capturing all relevant quality information about asource (this will provide a valuable resource for the INS). Structure metadata using an appropriate, agreed-upon metadata standard format is important.

e) Develop a good understanding of the data supplier, the context, and purpose of the data collection and the quality standards they uphold.

f) Build strong relationships with the data supplier, to ensure effective sharing of information – building a common understanding of each other's needs.

g) Put in place formal agreements, which outline clearly the INS and data supplier requirements, roles, andresponsibilities.

h) Carefully assess the value of acquiring and using an administrative source, against any risks and costs. Thiscan be with respect to the stability of a source over time and the risk of a data supplier failing to deliver dataon time or to the expected quality.

i) Ensure there is a sound legal basis to the receipt and use of an administrative source, with effective safeguards in place to protect the privacy of the data subjects.

j) Be clear and transparent about the use of administrative data, showing evidence that the benefits outweigh any privacy concerns.

k) Accept that objects, definitions, concepts, and time reference periods within an administrative source may not align with the statistical targets. It will, therefore, be necessary to transform data and make judgements on what levels of misalignment are acceptable.

l) Assess quality on a continuous basis (using the process and tools outlined) – responding to any anticipatedor known changes to a source. Document and publish the strengths and weaknesses associated with administrative sources, so that data users have confidence in the data and can take account of any limitations.

m) Be prepared that it will take time to understand and acquire administrative data sources, particularly, where work plan is required to develop registers.

List of appendices:

Appendix A. Quality check list for the Input phase – Source

3.0	**Quality check list for the discovery phase - Source**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyper-dimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future.

3.1	**Institutional environment**			
		Description	Score	Rank
3.1.1	**Data Provider Status (1)** Evaluate the overall risk that the data provider does not meet the quality requirements of the INS. Score: 1: high risk, 2: medium risk, 3: low risk, DK:don't know	Describe the type of organization (public, private, established or not, reputation, etc.).		
3.1.2	**Data Provider Status (2)**	Describe factors that could affect sustainability		

	Evaluate the risk that the ADS no longer be produced or made available by the data provider. Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know	through time (e.g., change of priorities, change of legislation, business instability, etc.).		
3.1.3	**Data Collector Status** If the data provider collects the data from other organizations, evaluate the additional risks broughtby these data collectors on both the supply and quality of the ADS (as described in 3.1.1 and 3.1.2). Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know, NA: Not applicable (no distinct data collectors)			

3.2	**Interpretability**			
		Description	Score	Rank
3.2.1	**Documentation** Is information available and accessible for the following? 1. File record layout 2. File data dictionary (objects and variables) 3. Reference period 4. Time period associated with the data (historical starting and end point) 5. Data collection procedures 6. Data treatment procedures (incl. capture, coding, editing, imputation) 7. Proportion of missing objects and variables 8. Imputation rate 9. Historical changes to the data collection and data treatment procedures 10. Other Give a score for each item listed above. Score: 1: no, 3: yes, DK: don't know	Describe globally the type of documentation and metadata available (concept, data collection, processing, etc.).		
3.2.2	**Feedback** Does the INS have the possibility to contact the data source provider to ask questions or to obtain			

	clarification about the information provided when deemed necessary? Score: 1: no, 2: maybe, 3: yes, DK: don't know			

3.3	**Accessibility**			
		Description	Score	Rank
3.3.1	**Restrictions** Assess the potential risk of facing restrictions in the access and use of the data by the INS. Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know	List the law, *Act*, or other legal or regulatory provision under which thedata source is being collected and maintained by the data collector and obtained by the data provider. Indicate any restrictions or limitations (can be stated as terms and conditions) regarding the use of the data by the INSand the dissemination of the resulting statistical outputs.		
3.3.2	**File transmission** Are the arrangements for the transmission of the data to the INS acceptable with respect toboth the security of the transmission and the equipment needed? Score: 1: no, 2: partially, 3: yes, DK: don't know	Describe the expected format (e.g. flat file, relational database, SAS/Sybase formats) andthe expected data transmission procedure. Specify if any specific software is required to have access to the data.		
3.3.3	**Public perception (1)** Will there be a need to carry out a privacy impact assessment or public consultations before the INS uses the data? Score: 1: yes, 2: maybe, 3: no, DK: don't know			
3.3.4	**Public perception (2)** If public consultations are likely to be necessary, how do you evaluate the risk that this will slow			

	down significantly the acquisition process or affect its relevance (for example, if access to the data is limited)? Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know, NA: Not applicable (public consultationsnot needed)			
3.3.5	**Ease of access** Are the data expected to be easily readable once transmitted to the INS? Score: 1: no, 2: partially, 3: yes, DK: don't know	Describe the level of efforts to read the data once transmitted.		

3.4	**Timeliness**			
		Description	Score	Rank
3.4.1	**File delivery (1)** Are the terms of delivery acceptable to the INS? For example, consider the timeliness for the availability of the data taking into account the INS requirements (e.g., production cycle). Score: 1: no, 2: partially, 3: yes, DK: don't know	Document the potential terms of file delivery (timing and frequency).		
3.4.2	**File delivery (2)** Estimate the risk for the INS that the data source is not delivered on time. Score: 1: high risk, 2: medium risk, 3: low risk, 0: don't know			

3.5	**Relevance**			
		Description	Score	Rank
3.5.1	**Confidentiality** Given the INS confidentiality policy, how do you evaluate the risk that it could limit the intended use of the data? Score: 1: high risk, 2: medium risk, 3: low risk, 0: don't know			

3.5.2	**Objects** Does the file contain the type of objects and object sets needed to meet the INS requirements for a given statistical output? Score: 1: no, 2: partially, 3: yes, 0: don't know	Describe globally the object types and object sets of the ADS.		
3.5.3	**Variables** Does the file contain the variables needed to meet the INS requirements for a given statistical output? Score: 1: no, 2: partially, 3: yes, 0: don't know	Describe the relevant variables found in the ADS		

3.6	Source review
3.6.1	When the information from the data source provider is incomplete, the data source collector should be contacted if different from the data source provider. Is there a need to contact the data source collector? If yes, approach the data source collector. Fill in missing information in the above checklist. Consider reflecting the additional costs related to this additional step in line 2.3.1 of the INS' requirements and intended use(s) of the data template.
3.6.2	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above. Looking at the summary table below, are important requirements met (i.e., high ranks do not have low scores or DK)? No: you may decide to stop the evaluation and report the findings;Yes: Continue with the METADATA checklist.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked						
TOTAL						

Appendix A1. Quality check list for the Input phase – Source filled in for the Source "Tax Agency"

3.0	**Quality check list for the discovery phase - Source**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyper-dimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future.

SOURCE: National Agency for Fiscal Authority
Compilers: Moldoveanu Ruxandra & Antoniade Ciprian Alexandru
Date:04-02-2022

3.1	**Institutional environment**			
		Description	Score	Rank
3.1.1	**Data Provider Status (1)** Evaluate the overall risk that the data provider does not meet the quality requirements of the INS. Score: 1: high risk, 2: medium risk, 3: low risk, DK:don't know	Describe the type of organization (public, private, established or not, reputation, etc.).	3	
3.1.2	**Data Provider Status (2)**	Describe factors that could affect sustainability		

	Evaluate the risk that the ADS no longer be produced or made available by the data provider. Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know	through time (e.g., change of priorities, change of legislation, business instability, etc.).	3	
3.1.3	**Data Collector Status** If the data provider collects the data from other organizations, evaluate the additional risks broughtby these data collectors on both the supply and quality of the ADS (as described in 3.1.1 and 3.1.2). Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know, NA: Not applicable (no distinct data collectors)		3	

3.2	**Interpretability**			
		Description	Score	Rank
3.2.1	**Documentation** Is information available and accessible for the following?	Describe globally the type of documentation and metadata available (concept, data collection, processing, etc.).		
	1. File record layout		3	
	2. File data dictionary (objects and variables)		3	
	3. Reference period		3	
	4. Time period associated with the data (historical starting and end point)			
	5. Data collection procedures			
	6. Data treatment procedures (incl. capture, coding, editing, imputation)	*we don't know which transformation they do at the end of the fiscal year for DEPOSIT	3	
	7. Proportion of missing objects and variables		2*	
	8. Imputation rate		3**	
	9. Historical changes to the data collection and data treatment procedures		N.A. -3	
	10. Other		3	
	Give a score for each item listed above. Score: 1: no, 3: yes, DK: don't know			
3.2.2	**Feedback** Does the INS have the possibility to contact the data source provider to ask questions or to obtain		3	

	clarification about the information provided when deemed necessary? Score: 1: no, 2: maybe, 3: yes, DK: don’t know			

3.3	**Accessibility**			
		Description	Score	Rank
3.3.1	**Restrictions** Assess the potential risk of facing restrictions in the access and use of the data by the INS. Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know	List the law, *Act*, or other legal or regulatory provision under which thedata source is being collected and maintained by the data collector and obtained by the data provider. Indicate any restrictions or limitations (can be stated as terms and conditions) regarding the use of the data by the INS and the dissemination of the resulting statistical outputs.	3* Privacy resolved with cryptation	
3.3.2	**File transmission** Are the arrangements for the transmission of the data to the INS acceptable by INS with respect toboth the security of the transmission and the equipment needed? Score: 1: no, 2: partially, 3: yes, DK: don’t know	Describe the expected format (e.g. flat file, relational database, SAS/Sybase formats) andthe expected data transmission procedure. Specify if any specific software is required to have access to the data.	3	
3.3.3	**Public perception (1)** Will there be a need to carry out a privacy impact assessment or public consultations before the INS uses the data? Score: 1: yes, 2: maybe, 3: no, DK: don’t know		3	
3.3.4	**Public perception (2)** If public consultations are likely to be necessary, how do you evaluate the risk that this will slow		N.A.	

	down significantly the acquisition process or affect its relevance (for example, if access to the data is limited)? Score: 1: high risk, 2: medium risk, 3: low risk, DK: don't know, NA: Not applicable (public consultationsnot needed)			
3.3.5	**Ease of access** Are the data expected to be easily readable once transmitted to the INS? Score: 1: no, 2: partially, 3: yes, DK: don't know	Describe the level of efforts to read the data once transmitted.	3	

3.4	**Timeliness**			
		Description	Score	Rank
3.4.1	**File delivery (1)** Are the terms of delivery acceptable to the INS? For example, consider the timeliness for the availability of the data taking into account the INS requirements (e.g., production cycle). Score: 1: no, 2: partially, 3: yes, DK: don't know	Document the potential terms of file delivery (timing and frequency).	3	
3.4.2	**File delivery (2)** Estimate the risk for the INS that the data source is not delivered on time. Score: 1: high risk, 2: medium risk, 3: low risk, 0: don't know		3	

3.5	**Relevance**			
		Description	Score	Rank
3.5.1	**Confidentiality** Given the INS confidentiality policy, how do you evaluate the risk that it could limit the intended use of the data? Score: 1: high risk, 2: medium risk, 3: low risk, 0: don't know		3	

3.5.2	**Objects** Does the file contain the type of objects and object sets needed to meet the INS requirements for a given statistical output? Score: 1: no, 2: partially, 3: yes, 0: don't know	Describe globally the object types and object sets of the ADS.	2* *depending on the output we focus on	
3.5.3	**Variables** Does the file contain the variables needed to meet the INS requirements for a given statistical output? Score: 1: no, 2: partially, 3: yes, 0: don't know	Describe the relevant variables found in the ADS	As above	

3.6	Source review
3.6.1	When the information from the data source provider is incomplete, the data source collector should be contacted if different from the data source provider. Is there a need to contact the data source collector? If yes, approach the data source collector. Fill in missing information in the above checklist. Consider reflecting the additional costs related to this additional step in line 2.3.1 of the 'INS requirements and intended use(s) of the data' template.
3.6.2	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above. Looking at the summary table below, are important requirements met (i.e., high ranks do not have low scores or DK)? No: you may decide to stop the evaluation and report the findings;Yes: Continue with the METADATA checklist.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked	18	2				
TOTAL						

Appendix B. Quality check list for the input phase – Metadata

4.0	**Quality check list for the discovery phase - Metadata**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyperdimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future. For the Metadata and Data checklists, an object oriented approach is used as it allows theapplication of the framework to any type of data. The component of the object oriented approach are the objects sets, which include all objects of the same type (i), to which areassociated (k) elements and (j) attributes or variables. For example, a data set containing individuals and businesses can be defined as containing two object sets, the first set is composed of individuals, $Object_{(1)}$ and the second set of businesses, $Object_{(2)}$. To the k-th element of $Object_{(1)}$, i.e., the k-th individual, are associated the variables (j).

4.1	**Interpretability (1), relevance (2)**			
		Description	Score	Rank
4.1.1	**Object sets and types (1)** Is information provided and sufficient to describe the objects sets and object types? 1: no, 2: partially, 3: yes	Describe the different object sets and object types.		
4.1.2	**Object sets and types (2)** Do the object sets and types suit the potential uses of interest? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.1 score=1)			
4.1.3	**Relational objects (1)** A relation between two object types can be regarded as a special kind of object type called a relational object. Identification keys are one type of relational objects that can allow linkage between objects. Is information provided and sufficient to describe the relational objects? 1: no, 2: partially, 3: yes, NA: Not applicable (no needfor relational objects in the potential uses of interest)	Describe the relational objects, if any.		
4.1.4	**Relational objects (2)** Do the relational objects suit the potential uses of interests? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.3 score =1), NA: Not applicable (no need for relational objects)			
4.1.5	**Object set coverage (1)** Is information provided and sufficient to describe the object set coverage? 1: no, 2: partially, 3: yes	Describe the object set coverage.		
4.1.6	**Object set coverage (2)** Does the object set coverage (for example, population units and statistical units, events) suit the			

	potential uses of interest? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.5 score=1)			
4.1.7	**Object set - time dimension - reference period (1)** Is information provided and sufficient to determine the reference period? 1: no, 2: partially, 3: yes	Describe the reference period.		
4.1.8	**Object set - time dimension - reference period (2)** Considering the potential uses, is the reference period adequate? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.7 score=1)			
4.1.9	**Object set - time dimension[1] - historical changes (1)** Is information provided and sufficient to determine changes over time affecting the definition of objects? 1: no, 2: partially, 3: yes, NA: Not applicable (e.g.,single point in time ADS)	Describe historical changes related to the definition of the object sets.		
4.1.10	**Object set - time dimension - historical changes (2)** If changes over time do occur, how do you rate the limitations on the potential uses? 1: high, 2: medium, 3: low, DK: don't know (to be used if 4.1.9 score=1), NA: Not applicable (e.g., sin-glepoint in time ADS)			
4.1.11	**Variables (1)** Is information provided and sufficient to describe the variables of interest? 1: no, 2: partially, 3: yes	Describe the variables (of interest).		
4.1.12	**Variables (2)** How close are the variables to those needed for the potential uses of interest? DK: Don't know (description missing or insuffi-cient)1: Not the same and conversion is impossible			

[1] The **Sensitivity dimension** is mainly used to determine the effect of time-dependent changes in the populationcomposition on data quality (Daas et al., 2008b).

	2: Not the same but conversion is possible 3: Identical			
4.1.13	**Variables - time dimension - historical changes (1)** Is information provided and sufficient to determine changes over time affecting the definition of variables? 1: no, 2: partially, 3: yes, NA: Not applicable (e.g.,single point in time variables)	Describe historical changes related to definition of the variables.		
4.1.14	**Variables - time dimension - historical changes (2)** If changes to variable do occur over time, how do you rate the limitations on the potential uses? 1: high, 2: medium, 3: low, DK: don't know (to be used if 4.1.13 score=1), NA: Not applicable (e.g.,single point in time ADS)			

4.2	**Coherence, accuracy**			
		Description	Score	Rank
4.2.1	**Variables - unique combination** Is there a combination of variables present that can be used to uniquely identify the objects within the object set? 1: no, 2: partially, 3: yes			
4.2.2	**Variables - collection procedures** Based on the information you have about the collection procedures (referring to 3.1.2, item 5), arethere important limitations that would affect the quality of the data? 1: yes, 2: maybe, 3: no, DK: don't know (to be used if 3.1.2 item 5 score = 1)	Describe the collection methods used. (Include frequency and timing of collection). Document if the collection vehicle (questionnaire or form) and the collection modehave been tested and if proxy response is accepted. Consider if there are legal or financial reasons why some fields of interest are likely to be very reliable.		

4.2.3	**Variables - data capture and coding** Are quality assurance processes in place at data capture? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe how the data are captured and coded, as well as how quality assurance is ensured during those stages.		
4.2.4	**Variables - editing** Are values that failed edits easily identifiable in the file? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe the methods used for edit checks (consistency edits, out-lier detection, etc.).		
4.2.5	**Variables- imputation** Are modified values easily identifiable in the file? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe the imputation methods.		
4.2.6	**Variables - treatment - changes over time** If important changes over time do occur for the var-ious treatment processes described (data collection, coding, editing and imputation), how doyou rate the limitations on the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know, NA: Notapplicable (e.g., no changes over time)			

4.3	**Metadata quality assessment summary**
4.3.1	When the information from the data source provider is incomplete, the data source collector should be contacted if different from the data source provider. Is there a need to contact the data source collector? If yes, approach the data source collector. Fill in missing information in the above checklist. Consider reflecting the additional costs related to this additional step in line 2.3.1 of the 'INS requirements and intended use(s) of the data' template.

4.3.2	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above. Looking at the summary table below, are important requirements met (i.e., high ranks do not have low scores or DK)? No: you may decide to stop the evaluation and report the findings;Yes: Continue with the DATA checklist.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked						
TOTAL						

Appendix B1. Quality check list for the input phase – Metadata filled in for the Source "Tax Agency"

4.0	**Quality check list for the discovery phase - Metadata**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyper-dimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future. For the Metadata and Data checklists, an object-oriented approach is used as it allows theapplication of the framework to any type of data. The component of the object-oriented approach are the objects sets, which include all objects of the same type (i), to which areassociated (k) elements and (j) attributes or variables. For example, a data set containing individuals and businesses can be defined as containing two object sets, the first set is composed of individuals, $Object_{(1)}$ and the second set of businesses, $Object_{(2)}$. To the k-th element of $Object_{(1)}$, i.e., the k-th individual, are associated the variables (j). Date: 18/02/2022 Compiler: Moldoveanu Ruxandra & Antoniade Ciprian Alexandru Source: Tax agency

4.1	**Interpretability (1), relevance (2)**			
4.1.1	**Object sets and types (1)** Is information provided and sufficient to describe the objects sets and object types? 1: no, 2: partially, 3: yes	Describe the different object sets and object types.	3	
4.1.2	**Object sets and types (2)** Do the object sets and types suit the potential uses of interest? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.1 score=1)	Intended usage: Update the frame in the inter-census period Already used: monthly by wage statistics/business statistics (to check)	3	
4.1.3	**Relational objects (1)** A relation between two object types can be regarded as a special kind of object type called a relational object. Identification keys are one type of relational objects that can allow linkage between objects. Is information provided and sufficient to describe the relational objects? 1: no, 2: partially, 3: yes, NA: Not applicable (no needfor relational objects in the potential uses of interest)	Describe the relational objects, if any. Relation: Employer - employee Relation: person - household NOT FROM THIS DATA ADDITIONAL CHECKS FOR HOUSEHOLD COMPOSITIONS	3	
4.1.4	**Relational objects (2)** Do the relational objects suit the potential uses of interests? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.3 score =1), NA: Not applicable (no need for relational objects)	The main use of this source will be updating counts for residents in Romania. For this we don't need relational objects. The source does not provide info on the relational objects for it additional usage as sample frame. FURTHER CONSIDERATIONS ARE NEEDED		
4.1.5	**Object set coverage (1)** Is information provided and sufficient to describe the object set coverage? 1: no, 2: partially, 3: yes	Describe the object set coverage.	3	
4.1.6	**Object set coverage (2)** Does the object set coverage (for example, population units and statistical units, events) suit the		3	

	potential uses of interest? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.5 score=1)			
4.1.7	**Object set - time dimension - reference period (1)** Is information provided and sufficient to determine the reference period? 1: no, 2: partially, 3: yes	Describe the reference period.	3	
4.1.8	**Object set - time dimension - reference period (2)** Considering the potential uses, is the reference period adequate? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 4.1.7 score=1)	Monthly data =income from salary Quarterly data =income from salary Yearly data = other income not from salary	3 3 2	
4.1.9	**Object set - time dimension[1] - historical changes (1)** Is information provided and sufficient to determine changes over time affecting the definition of objects? 1: no, 2: partially, 3: yes, NA: Not applicable (e.g.,single point in time ADS)	Describe historical changes related to the definition of the object sets. Needs to distinguish between changes in the data structures and changes in the objects	3	
4.1.10	**Object set - time dimension - historical changes (2)** If changes over time do occur, how do you rate the limitations on the potential uses? 1: high, 2: medium, 3: low, DK: don't know (to be used if 4.1.9 score=1), NA: Not applicable (e.g., sin-glepoint in time ADS)		3	
4.1.11	**Variables (1)** Is information provided and sufficient to describe the variables of interest? 1: no, 2: partially, 3: yes	Describe the variables (of interest).	3	
4.1.12	**Variables (2)** How close are the variables to those needed for the potential uses of interest? DK: Don't know (description missing or insuffi-cient)1: Not the same and conversion is impossible and elaborated analyses are needed 2: Not the same but conversion is possible 3: Identical	Purpose: derive resident status Purpose: derive net in-come/age Purpose: derive the gender	1 2 3	

[1] The **Sensitivity dimension** is mainly used to determine the effect of time-dependent changes in the populationcomposition on data quality (Daas et al., 2008b).

4.1.13	**Variables - time dimension - historical changes (1)** Is information provided and sufficient to determine changes over time affecting the definition of variables? 1: no, 2: partially, 3: yes, NA: Not applicable (e.g.,single point in time variables)	Describe historical changes related to definition of the variables.	3	
4.1.14	**Variables - time dimension - historical changes (2)** If changes to variable do occur over time, how do you rate the limitations on the potential uses? 1: high, 2: medium, 3: low, DK: don't know (to be used if 4.1.13 score=1), NA: Not applicable (e.g.,single point in time ADS)		3	

4.2	**Coherence, accuracy**			
		Description	Score	Rank
4.2.1	**Variables - unique combination** Is there a combination of variables present that can be used to uniquely identify the objects within the object set? 1: no, 2: partially, 3: yes		3 for people there is PIN 1 for HH	
4.2.2	**Variables - collection procedures** Based on the information you have about the collection procedures (referring to 3.1.2, item 5), arethere important limitations that would affect the quality of the data? 1: yes, 2: maybe, 3: no, DK: don't know (to be used if 3.1.2 item 5 score = 1)	Describe the collection methods used. (Include frequency and timing of collection). Document if the collection vehicle (questionnaire or form) and the collection modehave been tested and if proxy response is accepted. Consider if there are legal or financial reasons why some fields of interest are likely to be very reliable.	3 The collection tool is a form, that includes checks	

4.2.3	**Variables - data capture and coding** Are quality assurance processes in place at data capture? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe how the data are captured and coded, as well as how quality assurance is ensured during those stages.	3	
4.2.4	**Variables – editing** Are values that failed edits easily identifiable in the file? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe the methods used for edit checks (consistency edits, out-lier detection, etc.).	1	
4.2.5	**Variables- imputation** Are modified values easily identifiable in the file? 1: no, 2: partially, 3: yes, DK: don't know (to be usedif 3.1.2 item 6 = 1)	Describe the imputation methods.	1 Further check with col-leagues if this info are pro-vided by the Tax Agency	
4.2.6	**Variables - treatment - changes over time** If important changes over time do occur for the var-ious treatment processes described (data collection, coding, editing and imputation), how do you rate the limitations on the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know, NA: Notapplicable (e.g., no changes over time)		DK	

4.3	**Metadata quality assessment summary**
4.3.1	When the information from the data source provider is incomplete, the data source collector should be contacted if different from the data source provider. Is there a need to contact the data source collector? If yes, approach the data source collector. Fill in missing information in the above checklist. Consider reflecting the additional costs related to this additional step in line 2.3.1 of the 'INS requirements and intended use(s) of the data' template.

4.3.2	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above. Looking at the summary table below, are important requirements met (i.e., high ranks do not have low scores or DK)? No: you may decide to stop the evaluation and report the findings;Yes: Continue with the DATA checklist.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked						
TOTAL						

Appendix C. Metadata information template

The Appendix C is provided in Excel format *"6d app C admin-data-quality-metadata-info-template.xls"*, together with the report. An excerpt of sheet with Information and Instructions to use the templates is presented below.

Instructions for this metadata information template
This template is designed to capture the key aspects of quality for a given dataset in an organised way. Use it as part of a larger assessment of data quality, as described in the **Guide to reporting on administrative data quality** (see PDF in 'Available files').

You do not have to complete every single box in this template for it to be useful. Tailor the level of detail to the needs of your particular project or investigation. You should keep in mind that information you enter into this template will be very valuable to any other people who use the dataset in the future. Try to use easily understandable language, include all details and definitions, and do not assume too much of your readers.
The most important items you should completed first are:
General information: Items 1.1–1.6 including source agency, purpose of collection, summary of variables, and time span of the data.
Population: The target population, admin population, and reporting units. The items relating to coverage might not be possible to answer with a quick assessment but note anything you do know.
Variables: A short description of key variables. As work progresses, record the target concepts for the variables under investigation as they become known.
Collection: The timing/delay information and method of collection are important and should be easy to find out and record.
Note: You may be able to find much of this information for datasets already used at Statistics NZ in Colectica. All items will help you gain a sound understanding of a dataset's quality and the issues that might arise from using it for a different purpose. For example, understanding the original purpose of the data collection can guide you to which variables might be of higher quality than others, and to the likely coverage of the data.
Record any useful information for other questions but ignore any non-relevant boxes in the template. If you uncover relevant information later in the assessment, then add it – ideally the metadata information template for a given dataset should be improved and expanded as different people in Statistics NZ find out more about it. There should only be one metadata information template for everyone who uses a given dataset.

Published by Statistics New Zealand
xx March 2016
www.stats.govt.nz

Appendix D. Quality checklist for the Data acquisition, with identification and measurement of various types of errors

5.0	**Quality check list for the acquisition phase - Data**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyper-dimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future. For the Metadata and Data checklists, an object oriented approach is used as it allows theapplication of the framework to any type of data. The component of the object oriented approach are the objects sets, which include all objects of the same type (i), to which areassociated (k) elements and (j) attributes or variables. For example, a data set containingindividuals and businesses can be defined as containing two object sets, the first set is composed of individuals, $Object_{(1)}$ and the second set of businesses, $Object_{(2)}$. To the k-th element of $Object_{(1)}$, i.e., the k-th individual, are associated the variables (j).

5.1	**Accessibility**			
		Description	Score	Rank
5.1.1	**Readability** Can all the data in the ADS be accessed? 1: no, 2: partially, 3: yes, 0: don't know	Report the problems encountered.		

5.2	**Interpretability**			
		Description	Score	Rank
5.2.1	**Metadata compliance** Does the analysis of the data versus the metadata reveal anomalies that put the data into question and could limit the use of the data or require seeking clarifications from the data provider? 1: yes, 2: partially, 3: no, DK: don't know (absence orinsufficient metadata or data)	Describe the anomalies when the data do not comply with the metadata definition. For example, the format of the data is different than expected or the data contain values outside the expected range of values.		

5.3	**Accuracy, relevance**			
		Description	Score	Rank
5.3.1	**Object sets – over-coverage (1)** Duplicates cause over-coverage. Identification of duplicates is a simple process if unique identifiers are present on the file at the element level for a given object type. If it is not the case, a number of variables can be identified to serve as unique identifiers for the elements. When the file contains many records associated with a particular element (k), the identification of duplicates using this method may require more analysis as records that appear identical, when using	Describe the method used to identify duplicates, indicate the percentage of duplicatesand if they have been removed from the ADS.		

	a subset of variables, could rightfully be associated to the same k-th element for a given object type and not represent duplicates. Calculate the percentage of duplicate elements for each object type (i.e., within each object set). The percentage of duplicate elements by object type can be calculated as: $\frac{Nb\ of\ duplicate\ elements\ in\ the\ ADS}{Nb\ of\ elements\ in\ the\ ADS} \times 100\%$ Once duplicates are identified, **they should be removed from the ADS**. The cleaner file should be used in the subsequent steps of the quality evaluation process. Can duplicates be easily identified and removed? 1: no, 2: partially, 3: yes, DK: don't know (lack ofvariables to uniquely identified elements)			
5.3.2	**Object sets – over-coverage (2)** If a Reference Data Source (RDS) is available, it can be used to assess over-coverage, i.e., elements that are out of scope. Note that duplicate elements that could not be identified and removed from the ADS in 5.3.1 will also contribute to over-coverage. Identification of out of scope ADS elements can take place at the element level (micro level) through linkage with the RDS or at the aggregate level (macro level) for specific data space (for example, subpopulations) of interest. For example, you can calculate the percentage of over-coverage by object type: $\frac{Nb\ of\ ADS\ elements\ NOT\ in\ the\ RDS}{Nb\ of\ elements\ in\ the\ RDS} \times 100\%$ How do you rate the impact of over-coverage error given the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used to calculate over-coverage.		

5.3.3	**Object sets - under-coverage** If a Reference Data Source (RDS) is available, it can be used to assess under-coverage. The indicator should be calculated by object type (i.e., within each object set). For example, $$\frac{Nb\ of\ elements\ NOT\ in\ the\ ADS}{Nb\ of\ elements\ in\ the\ RDS} \times 100\%$$ Validation of the elements can take place at the element (micro level) through linkage with the RDS or at the aggregate level (macro level) for specific data space (for example, subpopulations) of interest. How do you rate the impact of over-coverage error given the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used to calculate under coverage		
5.3.4	**Object sets - selectivity (bias)** If a RDS is available, the selectivity indicators can beused to measure the degree to which in-scope elements included in the ADS differ from in-scope elements missing from the ADS (bias). Identify within the RDS elements that are common with the ADS (group A) and those only present on the RDS (group B). For group A and group B, compare summary statistics for the object type attributes (variables) within the data space of interest. Ideally, the variables used are in relation with the outcome or study variables of interest (histograms, bar plots can be used). Given the results obtained on the selectivity, how do you rate the limitations of the data given potential uses? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used.		

5.3.5	**Element - non-response** Element non-response occurs when all the data are missing for the variables of interest for a given element. Valid zeroes do not count as missing. Calculate the percentage of elements with all data missing for all variables of interest. This indicator canbe calculated separately for the main variables of interest. $\frac{Nb\ of\ ADS\ elements\ without\ data\ for\ all\ variables\ of\ interest}{Nb\ of\ ADS\ elements} \times 100\%$ Given the results obtained, how do you rate the impact of element non-response on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the ADS data available)			
5.3.6	**Object set - non-response - selectivity** By definition, the ADS non-responding elements have missing values for the main study variables but may contain information for other variables that are related with the main study variables. If it is the case, the selectivity indicators can be used to measure the degree to which the ADS responding elements differ from ADS non-responding elements. Ideally, the variables used are in relation with the outcome or study variables of interest. Compare summary statistics for the object type attributes available within the data space of interestfor responding and non-responding elements (histograms or bar plots can be used). Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.3.7	**Variables - non-response** Variable non-response occurs when values for specific variables of interest are missing for some			

	elements. Do not count valid zeroes as missing. When performing this analysis, the degree to which values are missing for a particular variable can be considered. For example, a variable could be considered has having missing values only when thedata are missing for a certain percentage of element(i.e., degree of missingness). Calculate the percentage of elements with data missing for specific variables included in the data space of interest. These should be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ missing\ values for\ variable\ i}{Nb\ of\ ADS\ elements\ in scope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.3.8	**Objects - variables non-response - selectivity** Assess if the objects with variable non-response have similar characteristics than objects with non-missing data. These should be calculated by object type (i.e., within each object set). Compare summary statistics for the objects attributes available within the data space of interestfor objects with variable non-response and objects without variable non-response. Ideally, the variablesused are in relation with the outcome or study variables of interest. Histograms or bar plots can be used. Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			

5.4	**Coherence**			
		Description	Score	Rank
5.4.1	**Variables - inconsistent values** Calculate the percentage of elements that violated edit rules for the range of acceptable values for a given variable of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ a\ valid\ value\ for\ variable\ i}{Nb\ of\ ADS\ elements\ in-scope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.4.2	**Variable set - coherence** Calculate the percentage of elements that violated edit rules for the coherence among variables. Only include variables that are part of the data space of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ objects\ with\ incoherent\ values}{Nb\ of\ ADS\ objects} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.4.3	**Variable - outliers** For variables of interest that do not have known values ranges, examine the distribution of the valuesto identify outliers. Consider using different outlier detection techniques such as the quartile method. Outliers should be flagged. These can also be calculated by object type (i.e., within each object set).			

	$\frac{Nb\ of\ ADS\ elements\ with\ outlier\ values\ for\ variable\ i}{Nb\ of\ ADS\ elements\ inscope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.4.4	**Variables - data processing adjustments by datasource provider** Calculate the percentage of elements with values adjusted (edited) by the data provider for the main variables of interest. These can also be calculated byobject type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ edited\ values\ for\ vriable\ i}{Nb\ of\ ADS\ elements\ in-scope\ for\ vaiable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the information available on edited variables)			
5.4.5	**Variables - imputation by data source provider** Calculate the percentage of elements with values imputed by the data provider for the main variables of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ imputed\ values\ for\ vriable\ i}{Nb\ of\ ADS\ elements\ inscope\ for\ vaiable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the information available on imputed variables)			
5.4.6	**Variables – rounding** Rounding can affect the value distribution.			

	Is there any evidence that rounding occurred for the main variables of interest? This can be detected by producing summary statistics and plots or histograms. Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., cannot be told by the information available)			

5.5	Linkages			
		Description	Score	Rank
5.5.1	**Relational objects – quality of linkage (1)** In many cases, linkage with Other Data Sources (ODS), such as a population frames, is a planned activity in order to fully exploit the ADS. For these cases, special considerations should be given to the quality of the variables used to perform the linkage. Note that other errors can also occur during the linkage process, however these types of errors are outside the scope of this checklist. 1. Calculate the percentage of elements: a) Elements linked in both the ADS and the ODS $\frac{Nb\ of\ ADS\ \ elements\ linked\ to\ ODS}{Nb\ of\ \ elements\ in\ the\ ADS} \times 100\%$ $\frac{Nb\ of\ ADS\ \ elements\ linked\ to\ ODS}{Nb\ of\ \ elements\ in\ the\ ODS} \times 100\%$ b) Percentage of elements linked unambiguously (strong link) c) Percentage of elements linked with a soft link (linking requirements that we lowered in order to link more elements) Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., cannot be told by the information available)			

5.5.2	**Relational objects – quality of linkage (2)** Calculate the percentage of elements not linked: d) ADS residual: ADS elements not linked $\frac{Nb\ of\ ADS\ elements\ not\ linked}{Nb\ of\ ADS\ elements} \times 100\%$ e) ODS residual: ODS elements not linked $\frac{Nb\ of\ ODS\ elelemnts\ not\ linked}{Nb\ of\ ODS\ objects} \times 100\%$ Large ODS residual is an indication of the limita-tionsregarding data integration. The ADS and ODS residuals may be an indication ofcoverage error in either the ADS or the ODS. Comparisons with the RDS, if available, can help resolve this. Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited im-pact,DK: don't know			

5.6	**Data quality assessment summary**
5.6.1	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked						
TOTAL						

5.7	**Final assessment and decision to acquire and use the ADS** A final decision towards data acquisition and use should take into account the quality assessments made in the four checklists (INS requirement, Source, Metadata and Data).
5.7.1	Explain on what basis the decision to acquire or not acquire the data is made and how the datais going to be used.

Appendix D1. Quality checklist for the Data acquisition, with identification and measurement of various types of errors filled in for the Source "Tax Agency".

5.0	**Quality check list for the acquisition phase - Data**
	We recommend performing the evaluation in three subsequent steps using: 1) the Source checklist, 2) the Metadata checklist and finally, 3) the Data checklist (a more in-depth quality assessment). Source, Metadata and Data can also be referred to as Hyper-dimensions. In general, a positive outcome in one step would indicate to pursue to the next step. The Source and Metadata checklists constitute the Discovery Phase. They can be performed prior to acquiring the data. In cases when the data are already available, the completion of the Source and Metadata checklists is also recommended as it contributes to document quality aspects that may be crucial for making a final decision regarding the fitness for use of the Administrative Data Source (ADS) given the intended use(s) by the INS. Each step covers a number of quality dimensions and associated quality indicators. A **score** must be assigned to each quality indicator to assess the ADS fitness for use with respect to the quality element. Special attention should be given when allocating the score as the scale is tailored to each question with a high score indicating a positive outcome. The INS can also prioritize the importance of each quality indicator being evaluated by assigning it a **rank** chosen between high, medium and low. The ranking can be used to identify a minimum set of quality indicators (the more important ones) that should be considered in the evaluation. The column "**Description**" is used to add relevant information and also to indicate when the information is not available. Note that when the INS is exploring the data with no specific intended use in mind and is doing so to rather discover potential uses, the evaluation still serves to assess the interpretability and completeness of the information provided. This information will be useful if potential uses for the same data file arise in the future. For the Metadata and Data checklists, an object oriented approach is used as it allows theapplication of the framework to any type of data. The component of the object oriented approach are the objects sets, which include all objects of the same type (i), to which areassociated (k) elements and (j) attributes or variables. For example, a data set containingindividuals and businesses can be defined as containing two object sets, the first set is composed of individuals, $Object_{(1)}$ and the second set of businesses, $Object_{(2)}$. To the k-th element of $Object_{(1)}$, i.e., the k-th individual, are associated the variables (j). Compilers: Ruxandra and Tiziana Date: 04/03/2022 Data: Tax Agency, last yearly provision (from 2021)

5.1	**Accessibility**			
		Description	Score	Rank
5.1.1	**Readability** Can all the data in the ADS be accessed? 1: no, 2: partially, 3: yes, 0: don't know	Report the problems encountered.	3	

5.2	**Interpretability**			
		Description	Score	Rank
5.2.1	**Metadata compliance** Does the analysis of the data versus the metadata reveal anomalies that put the data into question and could limit the use of the data or require seeking clarifications from the data provider? 1: yes, 2: partially, 3: no, DK: don't know (absence orinsufficient metadata or data)	Describe the anomalies when the data do not comply with the metadata definition. For example, the format of the data is different than expected or the data contain values outside the expected range of values.	3	

5.3	**Accuracy, relevance**			
		Description	Score	Rank
5.3.1	**Object sets - over-coverage (1)** Duplicates cause over-coverage. Identification of duplicates is a simple process if unique identifiers are present on the file at the element level for a given object type. If it is not the case, a number of variables can be identified to serve as unique identifiers for the elements. When the file contains many records associated with a particular element (k), the identification of duplicates using this method may require more analysis as records that appear identical, when using a subset of variables, could rightfully be associated to the same k-th element for a given object type andnot represent duplicates. Calculate the percentage of duplicate elements for each object type (i.e., within each object set).	Describe the method used to identify duplicates, indicate the percentage of duplicatesand if they have been removed from the ADS.	3	

	The percentage of duplicate elements by object type can be calculated as : $\frac{Nb\ of\ duplicate\ elements\ in\ the\ ADS}{Nb\ of\ elements\ in\ the\ ADS} \times 100\%$ Once duplicates are identified, **they should be removed from the ADS**. The cleaner file should be used in the subsequent steps of the quality evaluation process. Can duplicates be easily identified and removed? 1: no, 2: partially, 3: yes, DK: don't know (lack ofvariables to uniquely identified elements)			
5.3.2	**Object sets - over-coverage (2)** If a Reference Data Source (RDS) is available, it can be used to assess over-coverage, i.e., elements that are out of scope. Note that duplicate elements that could not be identified and removed from the ADS in 5.3.1 will also contribute to over-coverage. Identification of out of scope ADS elements can take place at the element level (micro level) through linkage with the RDS or at the aggregate level (macro level) for specific data space (for example, subpopulations) of interest. For example, you can calculate the percentage of over-coverage by object type: $\frac{Nb\ of\ ADS\ elements\ NOT\ in\ the\ RDS}{Nb\ of\ elements\ in\ the\ RDS} \times 100\%$ How do you rate the impact of over-coverage error given the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used to calculate over-coverage. Purpose: counting people resident RDS: pop resident list from Internal affairs	3	

5.3.3	**Object sets - under-coverage** If a Reference Data Source (RDS) is available, it can be used to assess under-coverage. The indicator should be calculated by object type (i.e., within each object set). For example, $$\frac{Nb\ of\ elements\ NOT\ in\ the\ ADS}{Nb\ of\ elements\ in\ the\ RDS} \times 100\%$$ Validation of the elements can take place at the element (micro level) through linkage with the RDS or at the aggregate level (macro level) for specific data space (for example, subpopulations) of interest. How do you rate the im~~pact of under-coverage er-~~ror given the potential uses of interest? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used to calculate under coverage The RDS is not easily available, we can use some proxy of it obtaining a raw percentage of under-coverage	1-2	
5.3.4	**Object sets - selectivity (bias)** If a RDS is available, the selectivity indicators can beused to measure the degree to which in-scope elements included in the ADS differ from in-scope elements missing from the ADS (bias). Identify within the RDS elements that are common with the ADS (group A) and those only present on the RDS (group B). For group A and group B, compare summary statistics for the object type attributes (variables) within the data space of interest. Ideally, the variables used are in relation with the outcome or study variables of interest (histograms, bar plots can be used). Given the results obtained on the selectivity, how do you rate the limitations of the data given potential uses? 1: high, 2: medium, 3: low, DK: don't know (e.g., areliable RDS is not available)	Include the results and a description of the methods used. More investigations are needed to evaluate the risk of bias	DK	

5.3.5	**Element - non-response** Element non-response occurs when all the data are missing for the variables of interest for a given element. Valid zeroes do not count as missing. Calculate the percentage of elements with all data missing for all variables of interest. This indicator canbe calculated separately for the main variables of interest. $\frac{Nb\ of\ ADS\ elements\ without\ data\ for\ all\ variables\ of\ interest}{Nb\ of\ ADS\ elements} \times 100\%$ Given the results obtained, how do you rate the impact of element non-response on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the ADS data available)		2	
5.3.6	**Object set - non-response - selectivity** By definition, the ADS non-responding elements have missing values for the main study variables but may contain information for other variables that are related with the main study variables. If it is the case, the selectivity indicators can be used to measure the degree to which the ADS responding elements differ from ADS non-responding elements. Ideally, the variables used are in relation with the outcome or study variables of interest. Compare summary statistics for the object type attributes available within the data space of interestfor responding and non-responding elements (histograms or bar plots can be used). Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)	Selectivity needs to be considered with more attention	DK	
5.3.7	**Variables - non-response** Variable non-response occurs when values for specific variables of interest are missing for some elements. Do not count valid zeroes as missing. When performing this analysis, the degree to which values are missing for a particular variable can be considered. For example, a variable could be considered has having missing values only when thedata are missing		2-3	

	for a certain percentage of element(i.e., degree of missingness). Calculate the percentage of elements with data missing for specific variables included in the data space of interest. These should be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ missing\ values for\ variable\ i}{Nb\ of\ ADS\ elements\ in scope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the variables available)			
5.3.8	**Objects - variables non-response - selectivity** Assess if the objects with variable non-response have similar characteristics than objects with non-missing data. These should be calculated by object type (i.e., within each object set). Compare summary statistics for the objects attributes available within the data space of interestfor objects with variable non-response and objects without variable non-response. Ideally, the variablesused are in relation with the outcome or study variables of interest. Histograms or bar plots can be used. Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the variables available)	As before		

5.4	**Coherence**			
		Description	Score	Rank
5.4.1	**Variables - inconsistent values** Calculate the percentage of elements that violated edit rules for the range of acceptable values for a given variable of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ a\ valid\ value\ for\ variable\ i}{Nb\ of\ ADS\ elements\ in-scope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the variables availabl		1	
5.4.2	**Variable set - coherence** Calculate the percentage of elements that violated edit rules for the coherence among variables. Only include variables that are part of the data space of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ objects\ with\ incoherent\ values}{Nb\ of\ ADS\ objects} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., unable to assess given the variables available)		DK	
5.4.3	**Variable - outliers** For variables of interest that do not have known values ranges, examine the distribution of the valuesto identify outliers. Consider using different outlier detection techniques such as the quartile method. Outliers should be flagged. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ outlier\ values\ for\ variable\ i}{Nb\ of\ ADS\ elements\ inscope\ for\ variable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest?		3	

	1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the variables available)			
5.4.4	**Variables - data processing adjustments by datasource provider** Calculate the percentage of elements with values adjusted (edited) by the data provider for the main variables of interest. These can also be calculated byobject type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ edited\ values\ for\ variable\ i}{Nb\ of\ ADS\ elements\ in-scope\ for\ vaiable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the information available on edited variables)		DK	
5.4.5	**Variables - imputation by data source provider** Calculate the percentage of elements with values imputed by the data provider for the main variables of interest. These can also be calculated by object type (i.e., within each object set). $\frac{Nb\ of\ ADS\ elements\ with\ imputed\ values\ for\ variable\ i}{Nb\ of\ ADS\ elements\ inscope\ for\ vaiable\ i} \times 100\%$ Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know (e.g., unable to assess given the information available on imputed variables)		DK	
5.4.6	**Variables – rounding** Rounding can affect the value distribution. Is there any evidence that rounding occurred for the main variables of interest? This can be detected by producing summary statistics and plots or histograms. Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., cannot be told by the information available)		3	

5.5	Linkability			
		Description	Score	Rank
5.5.1	**Relational objects – quality of linkage (1)** In many cases, linkage with Other Data Sources (ODS), such as a population frames, is a planned activity in order to fully exploit the ADS. For these cases, special considerations should be given to the quality of the variables used to perform the linkage. Note that other errors can also occur during the linkage process, however these types of errors are outside the scope of this checklist. 1. Calculate the percentage of elements : a) Elements linked in both the ADS and the ODS $\frac{Nb\ of\ ADS\ \ elements\ linked\ to\ ODS}{Nb\ of\ \ elements\ in\ the\ ADS} \times 100\%$ $\frac{Nb\ of\ ADS\ \ elements\ linked\ to\ ODS}{Nb\ of\ \ elements\ in\ the\ ODS} \times 100\%$ b) Percentage of elements linked unambiguously (strong link) c) Percentage of elements linked with a soft link (linking requirements that we lowered in order to link more elements) Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact, DK: don't know (e.g., cannot be told by the information available)			

5.5.2	**Relational objects – quality of linkage (2)** Calculate the percentage of elements not linked: d) ADS residual: ADS elements not linked $\frac{Nb\ of\ ADS\ elements\ not\ linked}{Nb\ of\ ADS\ elements} \times 100\%$ e) ODS residual: ODS elements not linked $\frac{Nb\ of\ ODS\ elelemnts\ not\ linked}{Nb\ of\ ODS\ objects} \times 100\%$ Large ODS residual is an indication of the limitationsregarding data integration. The ADS and ODS residuals may be an indication ofcoverage error in either the ADS or the ODS. Comparisons with the RDS, if available, can help resolve this. Given the results obtained, how do you rate the limitations of the data on the potential uses of interest? 1: high impact, 2: medium impact, 3: limited impact,DK: don't know			

5.6	**Data quality assessment summary**
5.6.1	Fill the summary table below with the number of quality indicators that obtained a given score and a given rank above.

Summary – Number of quality indicators given their score and rank.

Rank	Score					TOTAL
	3	2	1	DK	NA	
High						
Medium						
Low						
Not ranked						
TOTAL						

5.7	**Final assessment and decision to acquire and use the ADS** A final decision towards data acquisition and use should take into account the quality assessments made in the four checklists (INS requirement, Source, Metadata and Data).
5.7.1	Explain on what basis the decision to acquire or not acquire the data is made and how the datais going to be used.