# A Standard Tool for Metadata Quality Assessment of External Sources in the Statistical Production (QmetaTool)

E. Javier Orche Galindo[1]

[1] *National Statistical Institute of Spain (INE); javier.orche.galindo@ine.es*

**Abstract**
Knowledge and documentation of the source is vital to help us to use external data in the statistical production. A practical tool (QmetaTool) has been developed for these purposes and gathers the basic aspects to take into account in any use of external data in the statistical production. The issues are based in the state of clarity in the definition of units, time reference, geographical scope and variables (with their classifications and codifications). Clarity in the communications of changes in these definitions are evaluated too. Furthermore, the metadata assessment is focused on clarity in the description of checks and modifications in the database made by the owner. The last aspect is the state of knowledge in the existence of unique keys or a combination of variables that can be used to identify the populations units of microdata.
This standard tool (QmetaTool) offers the possibility to be used in any statistical operation fully or partial based on external sources. The time spent in the implementation of the tool is minimized and there is a graphical output for the overall situation of the metadata in the external data source. Just one assessment is needed by each source (no matter how many uses are made) and it can be implemented from any origin (public or private sector) and even in the use of big data for statistical purposes.

**Keywords:** quality, tool, metadata, assessment, microdata, external sources.

## 1. Introduction

Knowledge and documentation of the source is vital to help us to use external data in the statistical production. i.e.: how the data are collected, why they are gathered, how they are processed, concepts and definitions used, etcetera. A tool for a standard quality assessment of

these questions gave standard clues to a statistical office about strengths, weaknesses and what could be improved in the use of external sources for statistical purposes.

A practical tool (QmetaTool) has been developed for these purposes and gathers the basic aspects to take into account in any use of external source in the statistical production. The issues are based in the state of clarity in the definition of units, time reference, geographical scope and variables (with their classifications and codifications). Clarity in the communications of changes in these definitions are evaluated too. Furthermore, the metadata assessment is focused on clarity in the description of checks and modifications in the database made by the owner. The last aspect is the state of knowledge in the existence of unique keys or a combination of variables that can be used to identify the populations units.

This standard tool offers the possibility to be used in any statistical operation fully or partial based on external sources. The time spent in the implementation of the tool is minimized and there is a graphical output for the overall situation of the metadata in the external data source. Just one assessment is needed by each source (no matter how many uses are made) and it can be implemented from any origin (public or private sector) and even in the use of big data for statistical purposes.

The next chapter explains the main developments in the use of external sources with statistical purposes: preconditions to enable the use of them and methods to evaluate de usability of these data. The chapter three presents the features of the tool proposed (QnetaTool) for a standard metadata quality assessment, the section four explains the items used by the tool to assess the external source, the chapter five show the output obtained with the tool suggested and chapter six is about the conclusions and future actions after the use of the standard tool.

## 2. Quality developments in the use of external sources

In the last years, many statistical offices have wanted to increase the use of external microdata sources in the statistical production (basically through the use of administrative data for statistical purposes). For being enabled to use them, important European developments have been done in two main scopes:

*2.1. Preconditions to enable the use of external sources*

The first area is the foundation of a set of preconditions -in the light of the Nordic experiences- that enabled the use of administrative data sources for statistical purposes (UNECE, 2007). These are all necessary conditions and are as follows:

1) Legal foundation for the use of them
2) Public understanding and approval of the benefits of using them
3) Availability of one unified identification system for across the different sources used
4) Comprehensive and reliable systems of microdata sources
5) Cooperation of the microdata supplier

The conformance to these conditions will enable a statistical office to use external data for statistical purposes on a regular basis.

*2.2. Methods to evaluate the statistical usability of external sources*

The second scope is the development of methods to evaluate the statistical usability (i.e. quality) of external sources in a standardized way (Eurostat, 2003; BLUE-ETS, 2011). This basically involves answering to the following questions:

1) Should a data source be included in the production system of a statistical office?
2) How should it be used within the statistical production system?

As a data source may be used for other statistical purposes than being the primary source of data for statistics (e.g., sampling design, auxiliary data collection, imputation, etc.), the quality requirements on an external source depends on its potential role. This brings up that the quality of the external source has to be looked upon from two different views:

1) *From the view of consumer of statistics*. The consumer view concerns the quality of the final product, or the "Output quality" (Eurostat, 2009) and it is the quality way that has been traditionally looked upon.
2) *From the view of producer of statistics*. The producer view concerns to two areas:
    a. "Input data quality" – The preparations of the input needed for use in the production process (BLUE-ETS, 2011).

b. "Production process quality" – The gains in production efficiency of using the input (Laitila et al., 2011).

From the point of view of the "Input data quality" and according to the literature and practice, an instrument capable of determining this input quality of external source has to be efficient. It should not cost too much time and effort to determine the quality of the input because this could, theoretically, be determined every time a new delivery of the source or part of the source is received (Daas et al., 2010). It is therefore vital that the instrument developed focuses on the essential components of input quality. When the quality of a data source is determined two quality domains always need to be considered. These are quality in the metadata and in the data domain (Batini and Scannapieco, 2006). Metadata quality is not often studied independently of its Data counterpart but this approach has been successfully applied at Statistics Netherlands (Daas et al., 2008; 2010).

Statistics Netherlands has even developed a checklist for the determination of the quality components in the metadata domain (Daas et al., 2009). The major advantage of this approach is that the metadata quality components of a source can be determined independently of its content and, as a result, does not have to be checked every time the data in an source is studied (Daas et al., 2010). The quality indicators in metadata were grouped in four dimensions, namely:

1) Clarity
2) Comparability
3) Unique keys
4) Data treatment by the data source keeper

## 3. Features of the standard tool (QmetaTool)

The proposed tool is developed in a spreadsheet with an ".xlsx" format. Therefore, it can be opened by any of the usual applications for spreadsheets. There is no possibility to change the literals of the questions or modify their possible answers. Every item to assess has a literal question and a group box of option buttons with labels for each button. This groups related

controls into one visual unit in a rectangle with one label for each button and allows a single choice within a limited set of three or four mutually exclusive choices. An option button is also referred to as a radio button. Below the literal question and the group box there is a text box to insert an explanation according with some of the option buttons.

There are seven items to assess any external source and one item more if the external source are microdata. Finally, there is a text box to insert the main conclusions obtained and possible actions to do. These conclusions and actions can be type by the person performing the assessment or by other one.

## 4. Input items for the standard tool (QmetaTool)

### *4.1. Population units*

The first issue is based in the state of clarity in the definition of the population units. There are three possible answers depending on whether the description is clear, unclear or unknown. Only in the case of a clear description of the population units by the data supplier, this one has to be type in the text box below.

The scores of this item are as follows: 100% if the population units are defined clearly, 50% if the definition is unclear and 0% if the definition is unknown.

### *4.2. Time dimension*

This second question concerns to the clarity in the description of the period or point in time to which the data refer. There are also three possible answers depending on whether the description is clear, unclear or unknown. Only in the case of a clear description of the time interval of the data by the data supplier, this one has to be type in the text box below.

The scores of this item are as follows: 100% if the time dimension is defined clearly, 50% if the definition is unclear and 0% if the definition is unknown.

*4.3. Geographical scope*

The third concern is about clarity in the geographical scope to which the data refer to. The three possible answers are again: clear, unclear or unknown. Only in the case of a clear description of the geographical scope of the data by the data supplier, this one has to be type in the text box below.

The scores of this item are as follows: 100% if the geographical scope is defined clearly, 50% if the definition is unclear and 0% if the definition is unknown.

*4.4. Statistical variables*

The fourth issue is based in the state of clarity in the definition of the statistical variables (i.e., classifications. codifications used, etcetera). There are three possible answers depending on whether the description is clear, unclear or unknown. The list of main variables of statistical interest has to be type in the text box below.

The scores of this item are as follows: 100% if the statistical variables are defined clearly, 50% if the definition is unclear and 0% if the definition is unknown.

*4.5. Communication of changes*

This fifth question concerns if when the external source has adjusted a definition, this change is communicated clearly by the data supplier. There are now four possible answers depending on whether no changes have occurred or the changes have occurred and communication has been clear, unclear or it is suspected that changes have occurred but the communication has not occurred. Only in the case of a clear communication by the data supplier, the main changed definitions have to be described in the text box below.

The scores of this item are as follows: 100% if the communication has been clear or no changes are occurred, 50% if the communication has been unclear and 0% if the communication by the data supplier is suspected that not has occurred.

*4.6. Data checks by the supplier*

The sixth concern is about clarity in the description of data checks by the supplier: uniqueness of the units, plausibility of variables and combinations of variables, extreme values, etc. The three possible answers are: clear, unclear or unknown. Only in the case of a clear description of the data checks by the data supplier, the relevant supplier checks has to be described in the text box below.

The scores of this item are as follows: 100% if the data checks are described clearly, 50% if the description is unclear and 0% if the description is unknown.

*4.7. Data modifications by the supplier*

The seventh issue is about clarity in the description of data modifications by the supplier: editing and imputation data rules, default values for missing data, etcetera. There are now four possible answers depending on whether there are no data modifications or there are modifications and the description are clear, unclear or it is suspected that there are data modifications but the description of them is unknown. Only in the case of a clear description of the data modifications by the data supplier, the main supplier modifications has to be described in the text box below.

The scores of this item are as follows: 100% if there are no data modifications or the data modifications are described clearly, 50% if the description is unclear and 0% if the description is unknown but it is suspected that there are some modifications in the data.

*4.8. Identifying units: Unique keys or combinations of variables*

The eighth point is just in the case of microdata for record linkage and concerns about the existence of a unique key that can be used to identify population units. The three possible answers are: there is a unique key, there is not a unique key or it is unknown.

Also this issue is about the existence of a combination of variables that can be used to uniquely identify population units. The three possible answers are: it exists, it does not exists or it is unknown.

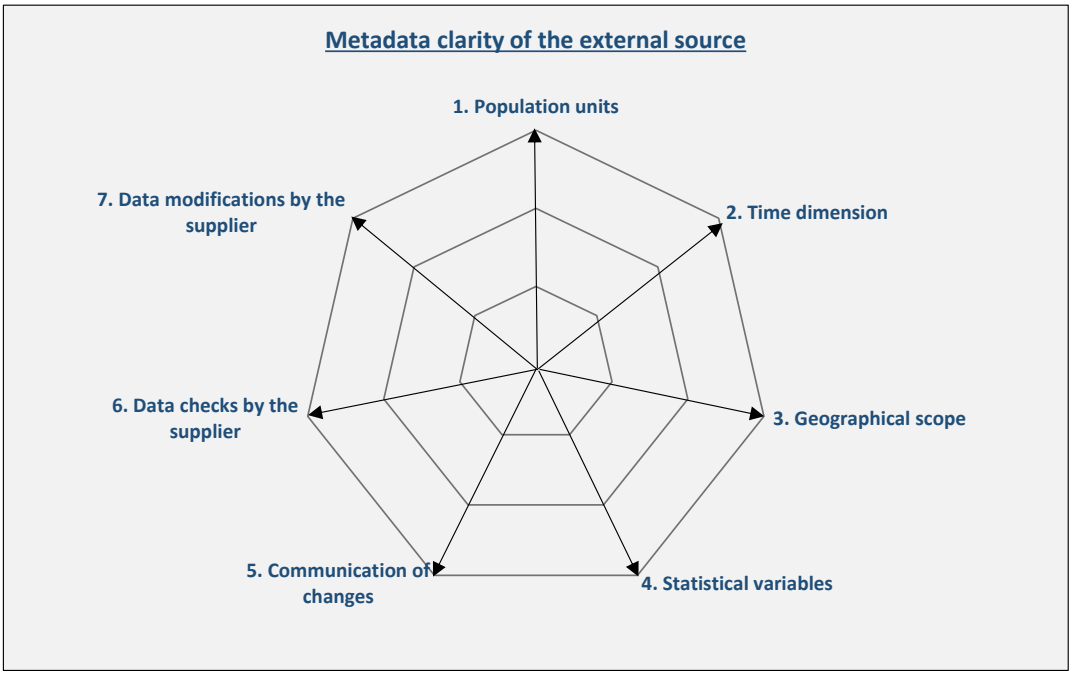
Only in the case of any of both exists, the main identification variables have to be described in the text box below. There is no scores for this item.

## 5. Outputs of the standard tool (QmetaTool)

The aim of the outcome is to have a global vision of clarity in the knowledge of external source metadata. When the assessment is made, the clarity of seven metadata items are quantified as 100% if it is "clear", 50% if it is "unclear" or 0% it is "missing" or "unknown". The tool shows the results of this scores and an average of these seven scores as an overall score.

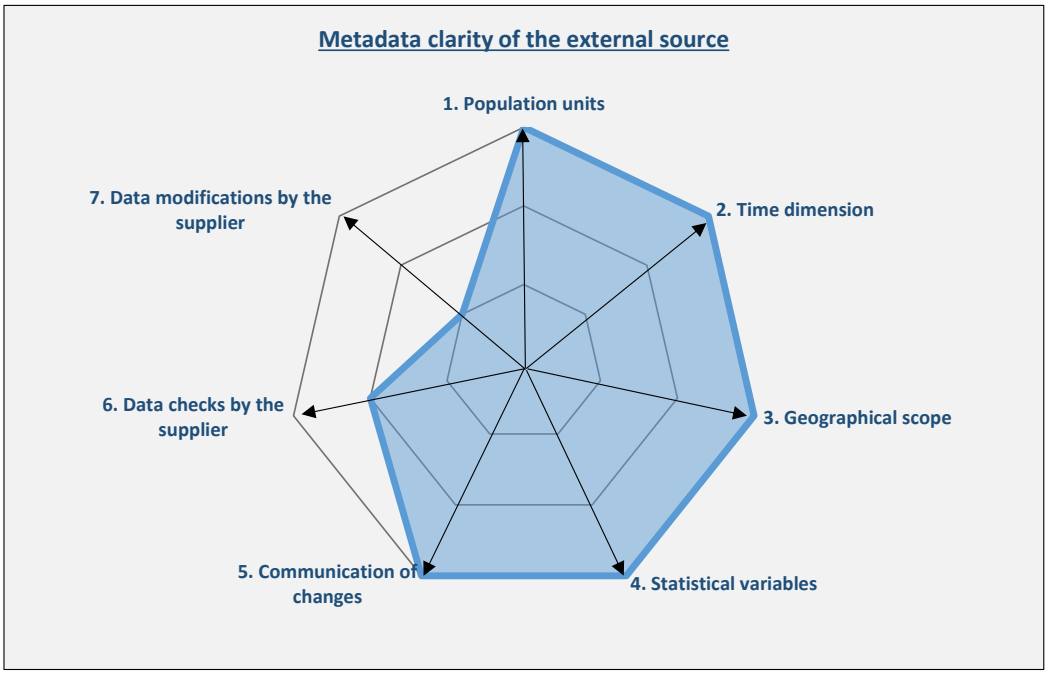
The global vision of the clarity in the knowledge of external source metadata is carried out by a radar chart, also known as a spider chart or a star chart because of its appearance, plots the values of each item along a separate axis that starts in the center of the chart and ends on the outer ring (see fig. 1). The inner ring matches to "missing" or "unknown" assessment, the middle ring is the "unclear" zone and the outer ring corresponds to "clear" in the item description, communication or definition.

**Figure 1. Radar chart (without the assessment)**



After the assessment of all seven items is made, there will be a filled blue area. This area will be as extensive as the clarity in metadata of the external source. The figure two shows a situation in the use of external sources where the first four or five items are all clear but there is a small knowledge about the sixth item and a complete ignorance about the seventh one.

**Figure 2. Radar chart (with the assessment)**




## 6. Outcomes of the standard tool (QmetaTool)

The global vision of the standard tool (QmetaToll) will give clues to the statistical office about strengths, weaknesses and what could be improved in the use of external sources for statistical purposes for each of them.

The conclusions and future actions after the use of the standard tool (QmetaTool) have to be aimed to clarify unknown or unclear metadata items. This would be done via the data supplier in the case that it was necessary to assess the potential use of an external source in the statistical production.

## 7. References

BLUE-Enterprise and Trade Statistics Project - Work Package 4.2 (2011). Report on methods preferred for the quality indicators of administrative data sources. Piet Daas, Saskia Ossen (CBS). With contributions of Martijn Tennekes (CBS), Li-Chun Zhang, Coen Hendriks, Kristin Foldal Haugen (SSB), Fulvia Cerroni, Grazia Di Bella (ISTAT), Thomas Laitila, Anders Wallgren, Britt Wallgren (SCB).

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008). Quality Framework for the Evaluation of Administrative Data. Paper for the Q2008 European Conference on Quality in Official Statistics. Rome, Italy.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., Arends-Toth, J. (2009). Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010). Determination of Administrative Data Quality: Recent results and new developments. Paper for the Q2010 European Conference on Quality in Official Statistics. Helsinki, Finland.

Eurostat (2003). Quality Assessment of Administrative Data for Statistical Purposes. Luxembourg.

Eurostat (2009). ESS Standard for Quality Reports. Methodologies and Working papers. Luxembourg.

Laitila, T., Wallgren, A., Wallgren B. (2011) Quality Assessment of Administrative Data, Statistics Sweden, Stockholm.

UNECE. (2007). Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics, Geneva: United Nations Publication.