BLUE-Enterprise and Trade Statistics

BLUE-ETS

www.blue-ets.eu

SP1-Cooperation-Collaborative Project
Small or medium-scale focused research project
FP7-SSH-2009-A
Grant Agreement Number 244767
SSH-CT-2010-244767

Deliverable 4.2

Dissemination level: PU

Report on methods preferred for the quality indicators of administrative data sources

Authors: 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)

Final version

28 September 2011

EUROPEAN COMMISSION
European Research Area

SEVENTH FRAMEWORK PROGRAMME

Funded under Socio-economic Sciences & Humanities

Deliverable 4.2

Report on methods preferred for the quality indicators of administrative data sources

Summary

This document contains an overview of the measurement methods developed for the quality indicators of administrative data when used as an input source for the statistical process of National Statistical Institutes. The indicators were identified and described in the first deliverable of workpackage 4 of the BLUE-ETS project and were grouped in the following five dimensions of quality: Technical checks, Accuracy, Completeness, Integrability, and Time-related dimensions.

The measurement methods described in this report form the basis of the quality-indicator instrument and software tool that will be developed for administrative data sources in the remainder of the project.

Index

1. INTRODUCTION

Many National Statistical Institutes (NSI's) want to increase the use of administrative data (i.e. registers) for statistical purposes. This requires that relevant administrative data sources need to be available in the home country of the NSI and that several preconditions have to be met in the home country of the NSI (Unece, 2007). The preconditions that enable an NSI to extensively make use of administrative sources in statistics production are: 1) legal foundation for the use of administrative data sources, 2) public understanding and approval of the benefits of using administrative sources for statistical purposes, 3) the availability of an unified identification system across the different sources used, 4) comprehensive and reliable systems in public administrations, and 5) cooperation among the administrative authorities. All of this will enable an NSI to use administrative data for statistics on a regular basis.

When the prerequisites described above are met, the statistical usability of administrative data sources becomes an important issue. To cope with fluctuations in the quality of these sources, it is essential that an NSI is able to determine its statistical usability (i.e. the quality) on a regular basis. This is an important issue because the collection and maintenance of an administrative data source are beyond the control of an NSI. It is the data source holder that manages these aspects. It is therefore of vital importance that an NSI has a procedure available that is able to determine the quality of administrative data - for statistical use – when it enters the office in a quick, straightforward, and standardised way. As yet, however, no standard instrument or procedure is available for such data sources (Daas et al., 2010). The development of such an instrument is the main focus of Workpackage 4 (WP4) of the BLUE-Enterprise and Trade Statistics (BLUE-ETS) project. An approach is needed that is practical, robust, efficient, and applicable to a whole range of administrative data sources.

1.1 First deliverable

In WP4 a first step was made in the development of an instrument that is able to quickly determine the quality of administrative data sources by identifying the quality ‘components’ that determine the input quality of these sources. Since a procedure was already available to study the metadata-quality of administrative data sources (Daas and Ossen, 2011), the focus of WP4 is on the determination of the quality of the data (Deliverable 4.1). The work started by identifying the quality ‘components’ essential for the input quality of administrative data. It was found that input quality was composed of five dimensions, namely: Technical checks, Accuracy, Completeness, Integrability, and Time-related dimensions (Annex A in Deliverable 4.1). For these dimensions a total of 28 unique indicators were identified. More details on the approach followed and the considerations made can be found in the first deliverable of WP4 (Deliverable 4.1). A review of the indicators proposed in this deliverable by experienced users at the various NSI’s involved in WP4 revealed that all indicators proposed were considered important. No indicators could be ignored.

1.2 Focus of this paper

The aim of this paper is to extend the work initiated in the first deliverable of WP4 by describing measurement methods for the quality indicators identified. Whenever a method is proposed that is included in the list of indicators of the ESSnet on Admin Data (ESSnet, 2010) or the standard list of quality indicators of Eurostat (2005) these will be referred to. The new methods developed affected a few of the indicators proposed in the first deliverable (see Annex A and B).

The consequences for implementing the measurement methods in a software tool are considered. In this document also an initial design for a “Quality Report Card” is proposed. This card will provide a short overview of the input quality of an administrative source.

The paper is organized as follows. Chapter 2 discusses what is actually meant by input quality of administrative sources. Recent insights are described and explained in this chapter. The quality indicators discussed in this document will be interpreted in the light of these different ‘views’. In chapter 3 for each indicator an overview is given of the measurement methods proposed and the consequences for implementing these methods in a software tool. In chapter 4 additional considerations for the software tool and a proposal for a Quality Report Card are discussed. The paper ends with chapter 5 that provides a preview of future work.

2. INPUT QUALITY DETERMINATION

2.1 On input quality

Input quality of administrative data can be looked upon from several points of view. One of them is a data archive point of view. In determining quality within this view the potential use of the data at the statistical office may be anticipated, but it is hardly or only to a very limited extent subject-specific. A different view on input quality will be taken by a statistical user of the data having already a specific use of the data in mind. This view is much more goal oriented. The deficiencies and strengths of the data are weighted accordingly: certain deficiencies of the data may not be important, while others are critical. Both points of view mentioned above are valid in a statistical context.

What is important to notice is that, even though it is the same data, the *assessment* of its quality may differ depending on the point of view taken. For example, suppose that an NSI obtains a source containing only data for a selective part of the population (e.g. the small and medium sized enterprises), but for which the data is of excellent quality. If the NSI plans to use this source on its own to produce Structural Business Statistics, then the data is obviously not good enough. However, if the source can be combined with, say, survey data collected for all the large enterprises, then the administrative data source would be highly useful. Now, the fact remains that the data has excellent quality for the sub-population it covers. From a data archiving point of view, it has good quality. But this may not necessarily be good enough for a specific statistical use as usefulness in this context depends not only on the source itself but also on the output requirements of the statistical process and the availability of additional data in the statistical system. Both views on input quality will be considered in this paper.

Note that in the second view the expected impact of the use of the data on *specific* output is considered. It is important to stress here that the work of WP4 focuses on the development of methods that are used at the moment a source enters the NSI (BLUE-ETS, 2010). This implies that the considerations of the quality of a data source on the impact on the output can, at that point in time, only be of an *implicit* and *preliminary* nature. Because of this nature the results of this particular quality assessment may differ from that made later on; after the source has actually been used in producing statistics[1]. In the remainder of this paper both views on input quality will be considered. In light of the considerations made above, the data archive point of view will be referred to as Data Source Quality (DSQ), while the more goal oriented view will be referred to as Input oriented Output Quality (IOQ).

2.2. On quality indicators

In this chapter the quality indicators proposed in the first deliverable of WP4 (Deliverable 4.1) are discussed from the viewpoint of DSQ and IOQ. Some changes have been made to the list of indicators originally presented in the first deliverable of WP4 (Deliverable 4.1). This is done to improve the structure and focus of the indicators proposed and is the result of considerations made during the development of measurement methods (see chapter 3). An updated version of the list of indicators is included as Annex A. Annex B lists all changes and summarises the motivations behind them.

2.2.1 Technical checks

This dimension predominantly consists of IT-related indicators for the data in a source. Apart from indicators related to the accessibility and correct conversion of the data, it also contains an indicator

[1] For completion, the reader is informed that process quality is not studied in WP4. For more information on that topic, the reader is referred to the findings of WP7 of the BLUE-ETS project. Their first results are described in the paper by Griffioen et al. (2011).

that checks if the data complies to the metadata-definition. The metadata can be a part of the delivery, either as a separate file or included as a header in the file (describing its content), but can also be provided by the data source holder in a separate process to the NSI.

The indicators included in the dimension Technical checks all focus on supporting the decision to either carry on using the data source or to report back to the data source holder. The latter decision could be the result of the fact that it is simply impossible to access the source or because it makes no sense to carry on. In other words, the indicators look at the necessary technical preconditions for quality assessment. It is expected that such checks are especially important when a new data source is being studied, but become less important once routine use has come into place. Although it will of course always be possible that a delivery suffers incidentally from technical problems. The end result of the Technical Checks dimension is essentially a go/no go decision. This clearly differentiates this dimension from the other dimensions identified for data (Deliverable 4.1). The indicators in this dimensions are both DSQ and IOQ oriented.

Table1. Scoring of the indicators in the Technical checks dimension on Data Source Quality (DSQ) and Input oriented Output Quality (IOQ).

Dimension indicators	Description	DSQ	IOQ
1. Technical checks	*Technical usability of the file and data in the file*		
1.1 Readability	Accessability of the file and data in the file	Yes	Yes
1.2 File declaration compliance	Compliance of the data in the file to the metadata agreements	Yes	Yes
1.3 Convertability	Conversion of the file to the NSI-standard format	Yes	Yes

2.2.2 Accuracy

The indicators in this dimension all originate from the sources of error that can occur when administrative data is used as input for NSI’s up to the point at which the data is linked to other (statistical) data sources. The indicators for objects point to the correctness of the units/events registered in the source, while the variable indicators focus on that of the values.

The quality indicators in this dimension not only focus on DSQ but also on IOQ. For example, the indicators Measurement error and Dubious values both refer to the correctness of the data. For DSQ these indicators can be measured with regard to the target concept of the source data, without having the potential statistical use in mind. However, for IOQ it is important to additionally consider the potential negative influence of these errors on the quality of the statistics produced.

Table 2. Scoring of the indicators in the Accuracy dimension on Data Source Quality (DSQ) and Input oriented Output Quality (IOQ).

Dimension indicators	Description	DSQ	IOQ
2. Accuracy	*The extent to which data are correct, reliable, and certified*		
2.1 Authenticity	Legitimacy of objects	Yes	Yes
2.2 Inconsistent objects	Extent of erroneous objects in source	Yes	Yes
2.3 Dubious objects	Presence of untrustworthy objects	Yes	Yes
2.4 Measurement error	Deviation of actual data value from ideal error-free measurements	Yes	Yes
2.5 Inconsistent values	Extent of inconsistent combinations of variable values	Yes	Yes
2.6 Dubious values	Presence of implausible values or combinations of values for variables	Yes	Yes

2.2.3 Completeness

For objects this dimension predominantly focuses on coverage issues, while the variable indicators are related to missing and imputed values. There are 4 object specific indicators and 2 indicators for variables in this dimension.

The indicators Undercoverage, Overcoverage, and Selectivity can be evaluated with regard to the target population of the source (i.e. DSQ), or potential target statistical populations (i.e. IOQ). In the latter case, one aims to assess whether the production process can deal with the coverage issues and what their anticipated influence on the output is. As mentioned before, it is important to realize that DSQ-related coverage and selectivity may affect the output quality no matter what the potential uses may be, and are therefore relevant for IOQ, albeit in a more implicit and preliminary way. Meanwhile, the indicator Redundancy examines the presence of multiple registrations of source-data objects. It is an indicator that explicitly refers to the quality of the delivered data, but contains often limited information on IOQ *per se*, since redundancy can in principle be removed. Of course, in cases where the multiply registered objects are associated with different variable values in

Table 3. Scoring of the indicators in the Completeness dimension on Data Source Quality (DSQ) and Input oriented Output Quality (IOQ).

Dimension indicators	Description	DSQ	IOQ
3. Completeness	*Degree to which a data source includes data describing the corresponding set of real-world objects and variables*		
3.1 Undercoverage	Absence of target objects (missing objects) in the source (or in business register)	Yes	Yes
3.2 Overcoverage	Presence of non-target objects in the source (or in business register)	Perhaps	Yes
3.3 Selectivity	Statistical coverage and representativity of objects	Yes	Yes
3.4 Redundancy	Presence of multiple registrations of objects	Yes	Perhaps
3.5 Missing values	Absence of values for (key) variables	Yes	Yes
3.6 Imputed values	Presence of values resulting from imputation actions by administrative data holder	Yes	Yes

addition, the effects on IOQ may not be negligible. Finally, the indicators Missing values and Imputed values are obviously relevant for both DSQ and IOQ, but possibly to a different extent, depending on how important the variables subjected to missing values are for the various outputs considered.

2.2.4 Time-related dimension

The quality indicators in this dimension are all related to time. The Timeliness, Punctuality, and Overall time lag indicators apply to the delivery of the input data file. The Delay indicator focuses on the recentness of the information stored. These indicators are all relevant for both DSQ and IOQ and their measurement methods are the same.

The remaining two other indicators are stability related. For objects a distinction is made between the dynamics that are intrinsic to the source-data population and the quality of updates. In other words, it is one thing that there are many births and deaths in the population from one period to the next, and it is another thing how quickly such changes are captured in the source. The former is always relevant for IOQ, but hardly a DSQ-issue, whereas the latter is relevant to both. The indicator Stability of variables is only meaningful for persistent units, but not for deaths and births. Still, the distinction between population intrinsic dynamics and the updating of them in the source applies here as well. In addition, it is an important IOQ-issue whether the variable composition covered by a source is stable over time or not, but this is not necessarily relevant for DSQ.

Table 4. Scoring of the indicators in the Time-related dimension on Data Source Quality (DSQ) and Input oriented Output Quality (IOQ).

Dimension indicators	Description	DSQ	IOQ
4. Time-related dimension	*Indicators that are time and/or stability related*		
4.1 Timeliness	Lapse of time between the end of the reference period and the moment of receipt of the data source	Yes	Yes
4.2 Punctuality	Possible time lag between the actual delivery date of the source and the date it should have been delivered	Yes	Yes
4.3 Overall time lag	Overall time difference between the end of the reference period in the source and the moment the NSI has concluded that it can definitely be used	Yes	Yes
4.4 Delay	Extent of delays in registration	Yes	Yes
4.5 Dynamics of Objects	Changes in the population of objects (new and dead objects) over time	No	Yes
4.6 Stability of Variables	Changes of variables or values over time	No	Yes

2.2.5 Integrability

This dimension contains indicators specific for the ease by which the data in the source can be integrated into the statistical production system of an NSI. The indicators for objects look at the comparability and ease of linking of the objects in the source to those commonly used by the NSI. The variable indicators either focus on the linking variable or compare the closeness of the values in the source to the facts of similar variables. A total of 4 indicators are included in this dimension.

Integrability is by definition related to the extent by which a source can be integrated in the production process. This implies that all indicators are relevant for IOQ. The extent to which an indicator is relevant for DSQ is in principle already be covered by the Accuracy, Completeness, and Time-related dimensions listed above.

Table 5. Scoring of the indicators in the Integrability dimension on Data Source Quality (DSQ) and Input oriented Output Quality (IOQ).

Dimension indicators	Description	DSQ	IOQ
5. Integrability	*Extent to which the data source is capable of undergoing integration or of being integrated*		
5.1 Comparability of Objects	Similarity of objects in source -at the proper level of detail - with the objects used by NSI	Perhaps*	Yes
5.2 Alignment of Objects	Linking-ability (align-ability) of objects in source with those of NSI	Perhaps*	Yes
5.3 Linking variable	Usefulness of linking variables (keys) in source	Perhaps*	Yes
5.4 Comparability of Variables	Proximity (closeness) of variables	Perhaps*	Yes

* Relevance to DSQ seems to essentially depend on earlier findings for indicators in the Accuracy, Completeness, and Time-related dimension. This would suggest that these are not relevant for DSQ.

3. MEASUREMENT METHODS FOR THE INDICATORS

During the meeting of WP4 in Stockholm (16-17 June) the ways to measure the indicators have been discussed. The essential considerations made for each indicator are:

- Can it be measured?
  a) If the answer is Yes:
    i) What measurement method(s) could be used?
    ii) What is the preferred measurement method?
    iii) What are the consequences of implementing the method(s) in software?

  b) If the answer is No:
    i) Can this information be provided by the data source holder?
      a) If the answer is Yes, this should be requested
      b) If the answer is No, this indicator can and should not be measured.

From the above it is clear that apart from the actual measurement, the implementation of the measurement methods in a tool also has to be taken into consideration. In the remainder of this chapter the indicators, proposed measurement methods, and considerations when implementing the methods in a tool are discussed for each dimension. An indicator that can not be measured should be removed from the list and also not be included in a tool. In the methods described in this chapter the use of percentages is preferred. However, the alternative of using absolute values is also a valid option. Whenever a method is proposed that is already included in the list of indicators developed for output by the ESSnet on Admin Data (ESSnet, 2010) or the standard list of quality indicators of Eurostat (2005) this will be referred to.

3.1 Technical checks

3.1.1 Readability

The readability indicator focuses on the accessibility of the file and the data in the file. Examples of problems in this area are a file of an unknown format, a corrupted file, a file with an unfamiliar character set, or a file that can not be decoded. Since a data source delivery can consist of a single (large) file and of small files containing records of single objects (or even corrections), both cases need to be dealt with. The main difference for accessibility is that it clearly refers to the accessibility of the data in the file in the first case, while it can also refer to the physical accessibility of the files in the second case. Therefore the following two methods of measurement are proposed.

Method 1: % of deliveries (or files) of the total deliveries with an unknown extension, that are corrupted, or cannot be opened
(for DSQ and IOQ).

Method 2: % of the total file which is unreadable (in size (MB/GB) or number of readable file records)
(for DSQ and IOQ).

Implementation:
For implementation it is essential that the user of the tool is able to select the file(s). An error message should be shown when the file(s) can not or can no longer be read. The tool must be able to read files in the various formats used by NSI's.

3.1.2 File declaration compliance

The file declaration compliance indicator considers the compliance of the data in the file to the metadata agreements. Examples of problems in this area are a file with a missing metadata description and a file with a lay-out that does not comply to the lay-out agreed upon. To enable determination of the compliance of the data to the metadata it is essential that the metadata is available. This could be either a part of the delivery or provided as a separate document.

Step 1: Determine whether metadata is available (Yes/No)
(for DSQ and IOQ).

If the answer to the first step is Yes, the next step needs to be applied.

Step 2: % of variables in the current delivery that differ from the metadata lay-out agreed upon in:

i) formats and names
ii) variable and attribute content
iii) categories defined for categorical variables
iv) ranges for numerical variables (if applicable, e.g. for age: 0-120)

(for DSQ and IOQ).

The second step resembles the approach commonly referred to as data profiling (Olson, 2003). In this technique analytical (IT-)techniques are used to discover inaccurate/erroneous data in files or databases.

Implementation:
To enable comparison with the metadata, the tool needs to be able to extract metadata-information from the data in the file. This is a procedure that is a common part of data profiling technology (Olson, 2003) for which commercial and open source software is available.

3.1.3 Convertability

The convertability indicator focuses on the conversion of the file to the NSI-standard format. Examples of problems in this area are file errors while decoding and corrupted data in the file after conversion. During conversion of a file or files errors can occur. This could result in a file that can not be opened or in a file with some errors for particular variables. These different errors need to be distinguished. Data sources composed of many small files are assumed to be converted to a single large file here.

Step 1: Can the file be opened after conversion (Yes/No)
(for DSQ and IOQ).

If the answer to the first step is Yes.

Step 2: % of objects with decoding errors or corrupted data
(for DSQ and IOQ).

In addition, step 2 of indicator 1.2 (paragraph 3.1.2) could be applied to determine the metadata compliance of the data *after* conversion.

Implementation:
Since it does not seem logical to perform the conversion to the NSI standard format in the tool, there are two options left. The first option is to use the error report (if any) generated by the conversion tool, while the other one is to rerun the measurement methods of indicator 1.2 to the

output of the conversion (paragraph 3.1.2). In both cases nothing new needs to be included in the tool.

3.1.4 Overall outcome

The overall outcome of the technical checks dimension is a go/no go decision. A delivery is either acceptable at a technical level or not; there is no in between. If a delivery is acceptable quality assessment should continue, if it is not the data source holder needs to be contacted in an attempt to solve the issue as quickly as possible. Looked upon in this way, the technical checks dimension can be considered a 'judgement portal' to the more statistical stage of administrative data quality evaluation.

3.2 Accuracy

The indicators in this dimension refer to either objects or to variables.

Accuracy of objects

3.2.1 Authenticity

The Authenticity indicator focuses on the legitimacy of objects in the source. This includes syntactic correctness of the identification key used (if present[2]) and the correspondence of the object in the source with the intended object in the real world. Examples of errors for this indicator are objects with an invalid (syntactically incorrect) identification key and objects with (syntactically correct but) a wrongly assigned identification key, respectively. The former objects can only be found when the numbers have to conform to certain syntactical defined structures and/or contain a check digit. For a data source with multiple identification keys, such as a Business register that for example contains an unique business identification code and an unique personal identification code for the owner, the syntactical correctness needs to be checked for all keys. For the identification of objects with wrongly assigned identification keys comparison with a reference list is needed. The term non-authentic objects is used for the objects identified by the methods proposed for this indicator.

Method 1: % of objects with a non-syntactically correct identification key
(for DSQ and IOQ).

Data sources that should only register objects once could -in principle- also be verified for multiple registrations of objects with the same key. This method is, however, included in the indicator redundancy (paragraph 3.3.4) and therefore not listed here.
Non-authentic object can be determined with the following method.

Method 2: % of objects for which the data source contains information contradictive to information in a *reference* list for those objects
(for DSQ and IOQ).

For DSQ, the reference list is the data source holder's master list of objects. For IOQ, the reference list is the target list of objects for the statistics under consideration. In both cases, objects with an identification key not included in the reference list should be ignored here (see paragraph 3.3.1).

Method 3: Contact the data source holder for their % of non-authentic objects in the source
(for DSQ and IOQ).

[2] If an unique key is not present or available for the objects under study the majority of the methods proposed in this document for which an unique key is required can not be used. An alternative might be to consider looking at unique combinations of values for a selected set of variables; such as sex, date of birth, and address (Arts et al., 2000).

Implementation:
For the first method the tool should not only enable users to select the key(s) that need to be verified but should also include a way to specify the rules used to check the correctness of the various keys. For the second method the user should be able to select the data source and a combination of variables. The same set of variables should also be selected from the reference list to which the data will be compared. Any differences need to be reported.

3.2.2 Inconsistent objects

The Inconsistent objects indicator looks at the overall consistency of the objects in the source. It focuses on the extent to which the objects in the source are not (or can not be made) internally consistent. The indicator is meaningful only when multiple objects are included in the data source. The indicator checks the inconsistency of objects within the data source; e.g. internal inconsistency. An example of such an error is a person in the data source that is assigned to multiple households or a local business unit that is assigned to more than one enterprise. In general, it can be stated that inconsistencies are errors that one is certain of.

Method 1: % of objects involved in non-logical relations with other (aggregates of) objects (for DSQ and IOQ).

Implementation:
The tool should be able to compare and check the relations between the objects in the source. This requires that logical relations need to be defined and that it should be possible to search for objects not satisfying these relations.

3.2.3 Dubious objects

The Dubious objects indicator focuses on the occurrence of untrustworthy objects. These are objects that are involved in relations with other objects that are presumably implausible but not necessarily incorrect (so-called soft rules). It is important to realize that -certainly for new data sources- it will be difficult to make these judgements or set these limits correct immediately for they require good knowledge of the data. This is obviously something that has to be build up over time.
The indicator is meaningful only when multiple objects are included in the data source. An example of a dubious object is a household to which 25 people are assigned with ages below 65.

Method 1: % of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects
(for DSQ and IOQ).

Implementation:
To implement this indicator in a software tool one needs to compare and check the relations between the objects in the source. This requires that implausible relations need to be defined and that it should be possible to search for objects not satisfying these relations.

Accuracy of variables

3.2.4 Measurement error

The Measurement error indicator looks at the correctness of the included value for a variable in the source. It checks to what extent the values actually measured by the data source holder correspond to the values that should be measured. There are several causes why a value lacks validity. For example: objects registered may have an interest in being registered in a particular way (i.e. reporting error), the administrative practice of the data source holder leads to biased entries (i.e. registration error), and the way the data is processed may lead to biased data (i.e. processing error; Bakker, 2010). Since all causes are related to the measurement process of the data source holder, it seems logical to contact the data source holder to obtain information on this indicator; when this is allowed of course. Only when the correctness (or incorrectness) of the values is marked by the data source holder, in the same or as part of a separate delivery, this is not needed. Independent of the cause of a measurement error, the effect for the NSI is always the same, i.e. an incorrect value in the data source.

Method 1: Only applicable when values not containing measurement errors are marked.
% of unmarked values in the data source for each variable
(for DSQ and IOQ).

The previous method can be used when the data source holder marks the correctness of the values for the variables in the source. When incorrect values are marked, this method can also be applied. In most cases these conditions will not be satisfied and the following method could be applied.

Method 2: Contact the data source holder and ask the following data quality management questions:

- Do they apply any design to the data collection process (if possible)?
- Do they use a process for checking values during the reporting phase?
- Do they use a benchmark for some variables?
- Do they use a checking process for data entry?
- Do they use any checks for correcting data during the processing or data maintenance?

(for DSQ and IOQ).

When it is found that the data source holder does not carry out data quality management or only very poorly, the NSI should seriously consider not using the data source or should ask the data source holder to satisfy a minimal set of requirements.
Apart from obtaining insight into the data collection method via method 2, the NSI should also analyse the data itself and check for wrong or suspicious values indicative for measurement errors. This can, for example be done, by looking for inconsistent or dubious values within the file (paragraph 3.2.5 and paragraph 3.2.6), by looking for identical or similar variables in other sources (paragraph 3.5.4), or by considering the dynamics of values for variables over time (paragraph 3.4.6).

Implementation:
Only the first method can be implemented in a software tool. This requires that the user must be able to specify how values are marked in the data file for a particular variable.

3.2.5 Inconsistent values

The Inconsistent values indicator looks at the consistency of values for combinations of variables in the source. It focuses on the extent to which the values for variables in the source are not (or can not be made) internally consistent. Again, it is important to realize that -certainly for new data sources- it will be difficult to make these judgements or set these limits correct immediately for they require good knowledge of the data. This is obviously something that has to be build up over time.
Examples of inconsistencies are a person in the data source that is male and pregnant or a person that is 10 years old and married. In general, it can be stated that inconsistent values are all values that do not satisfy relations that hold by definition; a combination of values that is not logical is erroneous.

Method 1: % of objects of which combinations of values for variables are involved in non-logical relations
(for DSQ and IOQ).

Implementation:
The tool should be able to compare and check the values of the variables for objects in the source. This requires that variables and objects can be selected and that logical relations need to be defined. The tool should enable the search for objects not satisfying these relations.

3.2.6 Dubious values

The Dubious values indicator checks for the occurrence of implausible combinations of values for the variables of an object. These are combinations of values that are suspicious but not necessarily incorrect. It is important to realize that -certainly for new data sources- it will be difficult to make these judgements or set these limits correct immediately for they require good knowledge of the data. This is obviously something that has to be build up over time.
An example of a dubious value is an enterprise for which the turnover per person employed is 10-times higher then the expected value usually found for similar types of enterprises. All objects with strange combinations of values for variables that can not be clearly explained are included in this indicator.

Method 1: % of objects with combinations of values for variables that are involved in implausible but not necessarily incorrect relations
(for DSQ and IOQ).

Implementation:
To implement this indicator in a software tool one needs to be able to compare and check the relations between the values of variables for objects in the source. This requires that variables and objects can be selected and that implausible relations need to be defined. The tool should enable the search for objects not satisfying these relations.

3.3 Completeness

The indicators in this dimension are assigned to either objects or variables.

Completeness of objects (Coverage)

3.3.1 Undercoverage

The Undercoverage indicator looks at the absence of target objects (missing objects) in the source. An example of this are objects active in the reference period covered by the source but not registered in it. To assess the undercoverage it is essential that at least some knowledge is available over the population as a whole. This could require the build up of a population of target objects from a whole range of data sources gathered over time. If this kind of information is not yet available, the NSI should start with that task first.

Method 1: % of objects of the *reference* list missing in the source
(for DSQ and IOQ).

For DSQ, the reference list to which the data source need to be compared is the master list; the population the administrative data holder has in mind. For IOQ, the reference list is the target list of the statistics under consideration. For DSQ, it might be needed to construct an administrative population from all previous data files obtained; a so-called expected population. One should be aware that it is difficult for some data sources to define the expected population, e.g. sources with data on international trade could miss objects because they simply did not export or import goods during the reporting period. For IOQ the Business register of the NSI often defines the target population for economical statistics. One has to realize that the scope of a data source could be limited to a specific branch only; e.g. an administrative data source with information on retail companies will clearly miss data on companies active in other branches. This can -but should not always- be interpreted strictly as undercoverage of that particular data source. A data source that suffers from administrative delay also has undercoverage problems when the source is immediately used after receipt.

Method 1 is included for indicator 12 in the list of quality indicators developed by the ESSnet on Admin. Data (ESSnet, 2010).

Implementation:
In a tool the user should be able to provide a reference file that defines all objects in the population under study (the master and target list respectively). This file will be used for comparison with the objects in the data source. To assess the overcoverage it is essential that at least some knowledge is available over the population as a whole. This could require the build up of a population of target objects from a whole range of data sources gathered over time. If this kind of information is not yet available, the NSI should start with that task first.

3.3.2 Overcoverage

The Overcoverage indicator focuses on the occurrence of non-target objects in the source. A data source containing data for objects that do not belong to the target population of the NSI (in that reference period) is an example of this.

Method 1: % of object in the source not *included* in the reference population
(for IOQ and perhaps for DSQ).

It can be expected that this method is more relevant for the NSI-target population; e.g. for IOQ. In this case the target list of the statistics under consideration should be used. It can however not be excluded that data source holder registers more objects than needed, which can only be detected using the master list of the data source holder.

This method is included for indicator 13 in the list of quality indicators developed by the ESSnet on Admin. Data (ESSnet, 2010) and for indicator A5 in the list of standard quality indicators of Eurostat (2005). The same is applicable to method 2 (see below).

From an administrative point of view, it is to be expected that objects reported in the source that were previously not reported by the data source holder are *new* objects. When the population register of the NSI is not updated frequently these objects need to be excluded from the calculation. To differentiate this, an additional method is defined.

Method 2: % of object in the source not *belonging* to the target population of the NSI
(for IOQ only).

Implementation:
In a tool the user should be able to provide a reference file that defines all objects in the population under study. This file will be used for comparison with the objects in the data source. For method 2 it is required that the user is able to indicate the objects he/she considers new.

3.3.3 Selectivity

The Selectivity indicator looks at the statistical coverage and representativity of objects in the source. Of course, to assess selectivity it is essential that knowledge is available over the population as a whole. This could require the build up of a population of target objects from a whole range of data sources gathered over time. If this kind of information is not yet available, the NSI should start with that task first.

The Selectivity indicator particularly focuses on objects that are not missing at random. This is for example the case when a data source contains information on a particular part of the population, such as large retail companies in the south of a country or on companies that report VAT in the first week after the end of a reporting period. Selectivity can be observed by statistical data inspection methods (Schulte Nordholt et al., 2011) and can also be calculated by the so-called Representativity indicator (R-indicator; Schouten et al., 2009).

Method 1: Use statistical data inspection methods, such as histograms, to compare a background variable (or more than one) for the objects in the data source and the *reference* population
(for DSQ and IOQ).

Examples of this approach can be found in Schulte Nordholt et al. (2011) and Templ and Alfons (2009). For DSQ the master list should be used as a reference, while for IOQ the target list should be used.

Method 2: Use of more advanced graphical methods, such as tableplots (Tennekes et al., 2011).
(for DSQ and IOQ).

Quantitative methods could be used as well (see Unwin et al., 2006; Part II).

Method 3: Calculate the R-indicator for the objects in the source (for DSQ and IOQ).

$$1 - 2\sqrt{\sum_{h=1}^{H} q_h (\lambda_{O,h} - \lambda_O)^2}$$

where h = stratum indicator
H = total number of strata
q_h = total number of records in stratum h / total number of records in the *reference* population
$\lambda_{O,h}$ = total number of records having a value in stratum h / total number of records in stratum *h*
λ_O = total number of records having a value / total number of records in the *reference* population

When all strata are covered equally by the data at hand $\lambda_{O,h}$ will be equal to λ_O for all strata. This results in a value of 1 for the R-indicator. The larger the difference between the coverage in strata the closer the value of the R-indicator will be to 0. For DSQ the master list of the data source holder should be used as the reference population, for IOQ the target list of the statistic under consideration should be used.

Implementation:
For the first method, depending on the plot type chosen, the user has to be able to select a plausible number of background variables. For a histogram and a frequency table one variable has to be selected. For a scatter plot and for a tableplot at least two variables are compared. An interesting example of how variable selection could be implemented in a graphical user interface can be found in the R-package VIM (Templ and Alfons, 2009). For the second method, a visualisation method could be implemented as described by Tennekes et al. (2011). For all methods the user has to be able to load a file in which the *reference* population of objects and additional background information on these objects is available. Also the user needs to state how the data under study should be grouped. To assure comparability over periods it would be preferable to store and include the way of grouping in a separate file. In using this approach care has to be taken of empty groups.

3.3.4 Redundancy

The Redundancy indicator looks at the occurrence of multiple registrations of identical objects in a source. In some data sources this should not occur.

Method 1: % of duplicate objects in the source (with the same identification number) (for DSQ only).

Method 2: % of duplicate objects in the source with the same values for a selection of variables (for DSQ and perhaps for IOQ).

Method 3: % of duplicate objects in the source with the same values for all variables (for DSQ).

Implementation:
For a tool it is essential that a user can specify which variables need to be included in the comparison.

Completeness of variables

3.3.5 Missing values

The Missing values indicator looks at the absence of values for variables in the data source. This can be looked upon from a single variable perspective or from a combination of variables. For this indicator missing values can be interpreted as the absence of a value or the presence of an indication that no value has been reported. To correctly assess this indicator, it is essential that good knowledge of the data has been build up over time.

Method 1: % of objects with a missing value for a particular variable (for DSQ and IOQ).

This method is similar to indicator A3 in the list of standard quality indicators of Eurostat (2005).

Method 2: % of objects with all values missing for a selected (limited) number of variables (for DSQ and IOQ).

This method should only focus on a limited number of variables and not all variables; method 2 is not getting at unit non-response.

Method 3: Use of graphical methods to inspect for missing values for variables (for DSQ and IOQ).

The user needs to be aware of the fact that for some objects it can be expected that no value is reported for one or more variables. The variables for which this is the case need to be excluded from the evaluation. Method 2 aims to deal with this issue. Method 3 uses graphical methods, such as tableplots (Tennekes et al., 2011), to provide a global overview of the data missing. The latter seems particularly suited for large datasets.

Implementation:
For the tool, it is essential that the user is able to specify what ‘value’ in the dataset is used to indicate a missing value. For instance, is the cell empty or do these cells all have the indication ‘no value’ or ‘Not a Number’ (NaN). When graphical methods are used, missing data should be

specified as an additional category for which a clearly distinguishable colour or shape is used (Tennekes et al., 2011).

3.3.6 Imputed values

The indicator Imputed values checks for the occurrence of values in the dataset resulting from imputation actions performed the data source keeper. An NSI is only able to determine this when the imputed cells are marked by the data source holder.

Method 1: % of imputed values per variable in the source
(for DSQ and IOQ).

This method is similar to indicator A4 in the list of standard quality indicators of Eurostat (2005) and resembles indicator 19 of the ESSnet on Admin. Data (ESSnet, 2010). When the data source holder does not mark imputed values, the alternative option is to contact the data source holder and request this information. The latter can only be done if the NSI is allowed to do this.

Method 2: Contact the data source holder and request the percentage of imputed values per variable
(for DSQ and IOQ).

Implementation:
If method 1 is implemented in a tool, the user should be able to specify how imputations are marked in the data file.

3.4 Time related dimension

The majority of the indicators in this dimension apply to both objects and variables. Objects and variables do have one separate indicator each.

3.4.1 Timeliness

The Timeliness indicator looks at the lapse of time between the end of the reference period and the moment of receipt of the data source. Data sources that can only deliver information after the time this information was needed in producing statistics are useless for the NSI. The first method reports the difference in days.

Method 1: Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports)
(for DSQ and IOQ)

Alternatively, the data of receipt could be replaced by the day the data source is available to the user of the data source at the NSI; this sort of indicates the point in time from whereon the user has access to the data.

Method 2: Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)
(for DSQ and IOQ).

Both methods are variants of indicator T2 and T3 in the list of standard quality indicators of Eurostat (2005).

Implementation:
It is questionable whether this difference needs to be calculated by a tool. If desired, a tool could of course calculate the difference (in days) between two dates provided by the user. However, it seems

more appropriate to let the tool provide the user with the first and last date of the entries in the data source. From thereon it is up to the user.

3.4.2 Punctuality

The Punctuality indicator focuses on the possible time lag between the actual delivery date of the source and the date it should have been delivered. This difference greatly affects the usability of the data source by the NSI. For statistics of high frequency, such as monthly statistics, the timely availability of administrative data is vital. Any delay in delivery affects the delicate balance between the timeliness of the output and the quality of the statistics produced.

Method 1: Time difference (days) = (Date of receipt by NSI) – (Date agreed upon; as laid down in the contract)
(for DSQ and IOQ).

The method is a variant of indicator T1 in the list of standard quality indicators of Eurostat (2005).

Implementation:
Here it is also questionable if this method should be implemented in the tool. For the sake of completeness a method could be implemented that provides the user the possibility to enter two dates after which it calculates the difference in days.

3.4.3 Overall time lag

The Overall time lag indicator looks at the overall time difference between the end of the reference period in the source and the moment the NSI has concluded that the source can definitely be used. In contrast to the previous indicators in this dimension, this indicator includes the predicted time needed for evaluation. This prompts the NSI to efficiently evaluate sources routinely which is especially important for statistics of high frequency.

Method 1: Total time difference (days) = (Predicted date at which the NSI declares that the source can be used) – (Date of the end of the reference period over which the data source reports)
(for DSQ and IOQ).

The method is a variant of indicators T2 and T3 in the list of standard quality indicators of Eurostat (2005).

Implementation:
It is questionable if this method should be implemented in the tool. For the sake of completeness a method could be implemented that provides the user the possibility to enter two dates after which it calculates the difference in days.

3.4.4. Delay

The Delay indicator looks at the speed at which changes are captured in registration; this applies both to changes in the population composition and the values of variables for objects in the population. Data sources with delays provide the NSI with -more or less- outdated data. The first method proposed involves contacting the data source holder to report their experiences

Method 1: Contact the data source holder to provide their information on registration delays
(for DSQ and IOQ).

This difference could be calculated without contacting the data source holder by comparing the registered facts and objects in various subsequent deliveries of an administrative data source (Zhang, 2009). The method proposed calculates the difference in days.

Method 2: Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)
(for DSQ and IOQ).

To calculate an overall delay for two or more variables the results could be averaged. Delays in capturing registration add additional time to the outcomes of the indicators 4.1, 4.2, and 4.3. The relevance for DSQ and especially IOQ largely depends on the relative size of the delay and the reporting period of the statistics concerned.

Implementation:
It is questionable if this method should be implemented in the tool. For the sake of completeness a method could be implemented that provides the user the possibility to enter two dates after which it calculates the difference in days.

Time-related dimension for objects

3.4.5 Dynamics of objects

The Dynamics of objects indicator looks at the usefulness of the source to identify changes in the population of objects over time. Changes are affected by the occurrence of new objects and the removal of objects no longer active. New objects are commonly referred to as births, inactive objects as deaths. Objects that remain a part of the population are referred to as persistent or alive. The adequate registration of births and deaths is especially important for data sources used to update the base register for objects (such as the population or business register). Any mistakes made in the update procedure greatly influence many (if not all) of the statistics that make use of these registers. Because this is so vital, in addition, a contact strategy (or survey) could be used to check the accuracy of the births and deaths reported. The effects resulting from delays in registration are included in the Delay indicator of paragraph 3.4.4.

Consider a data source that reports on objects at two (consecutive) moments in time (*t* and *t-1*). Evaluating the changes (and non-changes) of objects between these moments in time, will produce the following.

Births *t* :	Number of objects at *t* but not at *t-1*
Alive *t* :	Number of objects at *t* and *t-1*
Deaths *t* :	Number of objects at *t-1* but not at *t*
Total objects *t* :	Births *t* + Alive *t*
Total objects *t-1*:	Alive *t* + Deaths *t*

The following method calculates the percentage of new objects in the source at *t* compared to the total number of objects at *t*.

Method 1: % Births *t* = (Births *t* / Total objects *t*) x 100%
= (Births *t* / (Births *t* + Alive *t*)) x 100%
(for IOQ only).

Method 2 is used to determine the percentage of deceased objects in the source at *t* compared to the total number of objects at *t*.

Method 2: % Deaths *t* = (Deaths *t* / Total objects *t*) x 100%

= (Deaths *t* / (Births *t* + Alive *t*)) x 100%
(for IOQ only).

Method 3 is identical to method 2 except for the fact that the number of deceased objects is now compared to the total number of objects at *t*-1. Missing objects are ignored here.

Method 3: % Deaths *t-1* = (Deaths *t* / Total objects *t-1*) x 100%
= (Deaths *t* / (Alive *t* + Deaths *t*)) x 100%
(for IOQ only).

Implementation:
To implement these indicators in a tool, the user must be able to load two data files, one for each period, and compare the objects in both data sources to one another. The objects in both sources need to be related preferably by an unique identification key.

Time-related dimension for variables

3.4.6 Stability of variables
The Stability of variables indicator focuses on the changes of variables or values over time. The indicator is only meaningful for persistent objects. Data sources that change the variable composition between subsequent deliveries may have problems in this area. For some variables it is essential that the values (such as the NACE code) remain reasonably stable between deliverables. This stability also depends on the level at which the values are observed. Both graphical techniques and calculations can be used to observe and express the stability of variable values. In this context it is important to mention that it depends on the variable whether frequent changes in values are a problem and whether this indicator should be applied.

Method 1: Use statistical data inspection methods to compare the values of specific variables for persistent objects in different deliveries of the source. Graphical methods that can be used are a bar plot and a scatter plot
(for IOQ only).

Method 2: % of Changes = (Number of objects with a changed value / total number of persistent objects with a value filled in for the variable under study) x 100%
(for IOQ only).

Method 3: A correlation statistical method can be used to determine to which extent values changed in the same direction for different object. For categorical data a method such as Cramers V can be used (Cramér, 1946)
(for IOQ only).

The level at which the changes are compared, e.g. micro- or aggregate level, will -very likely- greatly affect how the stability of the variable is observed. For instance, comparing NACE-code changes between data source deliveries at the 2 or 4-digit level will probably produce entirely different results for the same sources.

Implementation:
The tool should enable the user to select two data sources, containing data at *t* and *t*-1, and a variable that should be compared. In addition, the level of detail (or grouping) needs to be specified.

3.5 Integrability

The Integrability dimension looks at the ease by which the data in the source can be integrated in the existing statistical process. Indicators for object and variables are both included. All indicators are clearly relevant for IOQ. It is however questionable if they are relevant for DSQ. It is very likely that the quality components in this dimension could be attributed to components included in the Accuracy, Completeness, and/or Time-related dimensions. This reasoning is also supported by the fact that upon integration of a data source, a particular use is assumed, which points to IOQ and not to DSQ. The results of *preliminary* findings also point to IOQ.

Integrability of objects

3.5.1 Comparability of objects

The Comparability of objects indicator looks at the similarity of objects in the source -at the proper level of detail- with the objects used by NSI. It is important to mention that this indicator is evaluated before the data are *actually* integrated. A data source may report data for objects that differ from those used by the NSI and splitting up or converting them is difficult. Examples are sources containing data of local units, VAT units, families, and workers (to name a few). Comparability of objects is the first step in investigating the correspondence of the objects in the data source to those used by the NSI. Comparability can be measured at the micro-level, by:

Method 1: % of identical objects = (Number of objects with exactly the same unit of analysis and same concept definition as those used by NSI / Total number of relevant objects in source) x 100
(certainly for IOQ).

It is also possible that the objects included in the source are not identical to the objects used by the NSI but can be converted, for example, by combining units. The next method deals with this situation.

Method 2: % of corresponding objects = (Number of objects that, after harmonization, would correspond to the unit needed by NSI / Total number of relevant objects in source) x 100
(certainly for IOQ).

If needed, an additional distinction could be made between objects that are involved in a 1:1, 1:n, n:1, and n:m relation with an NSI-object(s) after harmonization.

Method 3: % of incomparable objects = (Number of objects that, even after harmonization, will not be comparable to one of the units needed by NSI / Total number of relevant objects in source) x 100
(certainly for IOQ).

Apart from micro-level similarity one may also check for similarities at the macro- or meso-level. A way to measure this is:

Method 4: % of non-corresponding aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 – fraction of objects of interest at the same aggregated level in source 2) x 100
(certainly for IOQ).

Of course the quality of the harmonization is also an issue. However, as mentioned before (Deliverable 4.1) input quality measurement certainly stops at that point.

Implementation:
Method 1 could be implemented in the same way as described under overcoverage (paragraph 3.3.2). Implementation of methods 2 and 3 is possible when the user is able to load a file in which the relations between the various identification numbers of the objects (used by the data source holder and the NSI) are laid down. When the tool needs to harmonize the objects according to the rules of the user, this is even more difficult to achieve. An alternative for this approach would be to allow the tool access to a dedicated conversion tool used by the NSI. Method 4 can only be implemented if in addition to the needs described for methods 2 and 3 an aggregation and selection function is included in the tool.

3.5.2 Alignment of objects

The Alignment indicator focuses on the linking-ability (align-ability) of the objects in the source with those of the NSI. An example of this is the degree of matching of objects in the population or business register (or objects in any of the other base registers) of the NSI to those in the source. In alignment the NSI objects are taken as the starting point. At the micro-level this can be measured by:

Method 1: % of identical aligned objects = (Number of objects in the business register with exactly the same unit of analysis and same concept definition as those in the source / Total number of relevant objects in business register) x 100
(certainly for IOQ).

The next method calculates the coverage of the (NSI-)objects in the source from the business register point of view.

Method 2: % of corresponding aligned objects = (Number of objects in the business registers that, after harmonization, correspond to units or parts of units in the source / Total number of relevant objects in business register) x 100
(certainly for IOQ).

If needed, an additional distinction could be made between objects in the business registers aligned via a 1:1, 1:n, n:1, and n:m relation with objects in the source.

Method 3: % of non-aligned objects = (Number of objects in the business register that, even after harmonization of the objects in the source, can not be aligned to one of the units in the source / Total number of relevant objects in business register) x 100
(certainly for IOQ).

The above method determines which part of the business objects under study are not covered by the source.

Apart from micro-level similarity one may also check for similarities at the macro- or meso-level. A way to measure this is:

Method 4: % of non-aligned aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 that can not be aligned + fraction of objects of interest at the same aggregated level in source 2 that can not be aligned) x 100
(certainly for IOQ).

Implementation:

For method 1 it is required that the user is able to load the objects in the business registers (their identification numbers) and the objects in the source. The implementation of this method will very much resemble the Overcoverage methods (paragraph 3.3.2) and the Comparability of objects methods (paragraph 3.5.1). The tool needs to support the fact that a user wants to focus on a particular part of the population in the business register, such as retail or industry. For methods 2 and 3 the user needs to additionally load a file in which the relations between the various identification numbers of the objects (used by the data source holder and the NSI) are laid down. When the tool needs to harmonize the objects according to the rules of the user, this is even more difficult to achieve. An alternative for this approach would be to allow the tool access to a dedicated conversion tool used by the NSI. Method 4 can only be implemented if in addition to the needs described for methods 2 and 3 an aggregation and selection function is included in the tool.

Integrability of variables

3.5.3 Linking variable

The Linking variable indicator looks at the usefulness of the linking variables (such as identification numbers) for the objects in the source. Examples of problems in this area are objects without linking variables and objects with linking variables that differ from those used by NSI (foreign keys).

Method 1: % of objects with no linking variable = (Number of objects in source without a linking variable / Total number of objects in the source) x 100
(certainly for IOQ).

Method 2: % of objects with (a) linking variable(s) different from the one(s) used by NSI = (Number of object in source with (a) linking(s) variable different from the one used by the NSI / Total number of objects with (a) linking variable(s) in the source) x 100
(certainly for IOQ).

For data sources that use a foreign key the result of method 2 will be 100%.

Methods 3: % of objects with correctly convertible linking variable(s) = (Number of objects in the source for which the original linking variable can be converted to one used by the NSI / Total number of objects with a linking variable in the source) x 100
(certainly for IOQ).

The methods listed above are related to the methods for the indicator Authenticity (paragraph 3.2.1; which contains a method that checks the syntactical correctness of the linking variable used) and those for Redundancy (paragraph 3.3.4; which checks for multiple registrations of objects). This suggests that there is, for DSQ, indeed no need to measure this indicator.

Implementation:
For method 1 it is required that the user is able to load the source and select the linking variable(s) in the source. For methods 2 and 3 the user needs to additionally load a file with all relevant linking variables used by the NSI. For method 3 the conversion rules or a file in which the relation is laid down between the different linking variables has to be loaded as well. The latter could be complicated.

3.5.4 Comparability of variables

The indicator Comparability of variables looks at the proximity (closeness) of the values of the variables in the source with those reported in other sources (such as administrative sources or sample surveys) used by the NSI. Both graphical techniques and calculations can be used to observe and express comparability of values. The methods for the indicators should preferably be applied to groupings of numeric and categorical variables. Some of the methods closely resemble the methods for the indicator stability of variables (paragraph 3.4.6). Main difference is that different sources are compared and not different periods of the same source. Comparison could occur at a grouped- and at micro-level.

Method 1: Use statistical data inspection methods to compare the totals of groupings of specific objects for variables in both sources. Graphical methods that can be used are a bar plot and a scatter plot. Distributions of values can also be compared.
(certainly for IOQ).

Method 2: The Mean Absolute Percentage Error (MAPE) that measures the mean of the absolute percentage error. MAPE has a lower bound of zero but has no upper bound. Alternatively the symmetric MAPE could be used. This method measures the symmetric mean of the absolute percentage error were the deviation between the percentage distributions is divided by the half-sum of the deviations.
(certainly for IOQ).

Method 3: A method derived from the chi-square test that evaluates the distributions of the numeric values in both data sets. For categorical data Cramers V (Cramér, 1946) could be used.
(certainly for IOQ).

Method 4: % of objects with identical variable values = (Number of objects in source 1 and 2 with exactly the same value for the variable under study / Total number of relevant objects in both sources) x 100
(certainly for IOQ).

Only when used in preliminary studies for data quality, the methods listed above could be relevant for DSQ. In all other cases it is only relevant for IOQ.

Implementation:
The tool should enable the user to select two data sources and a variable that should be compared. In addition, the level of detail (or grouping) needs to be specified. For comparison at the micro-level objects need to be linked.

4. ADDITIONAL CONSIDERATIONS

Apart from developing measurement methods, it is also important that WP4 -at this point in time- takes a look at the remaining future deliverables. These are a software tool, in which (part of) the measurement methods are implemented, and a Quality Report Card. Both are discussed in this chapter.

4.1 The software tool

The software tool is proposed to consist of the following four modules: an input data description module, an input data transformation module, the data quality measurement module, and the measurement output module (see Annex C).

I) The input data description module should be suited for administrative data and probably also for NSI survey data. This is needed because some measurement methods may require data from at least two sources for comparison. Important considerations here are: i) which file types should the tool support (only the small set of most used formats like flat files, DIF, CSV, XML, or more?), ii) the variable types supported (Character, Numeric), iii) the composition of the values shown (size, number of decimal places), iv) how the tool deals with empty values (NULL, NaN), and v) various variable roles (such as, identifier, e.g. primary key, and classification items).

II) The input data transformation module should be able to read input data and create workspace files for the data quality measurement methods module. Important considerations here are: i) the selection of input file/files, ii) the selection of variables from input file/files, and iii) the transformation of input to workspace files. Here, the question is raised if the conditional transfer of records (by defining filters for creating subsets of input files) needs to be included.

III) The data quality measurement methods module needs to implement all methods that can be measured for the indicators (see above). The module should contain the option to select for DSQ or for IOQ. Implementation of the measurement methods in the open source statistical software package R is preferred. Important considerations for this module are: i) the selection of a method, and ii) the selection of variables to which the method should be applied.

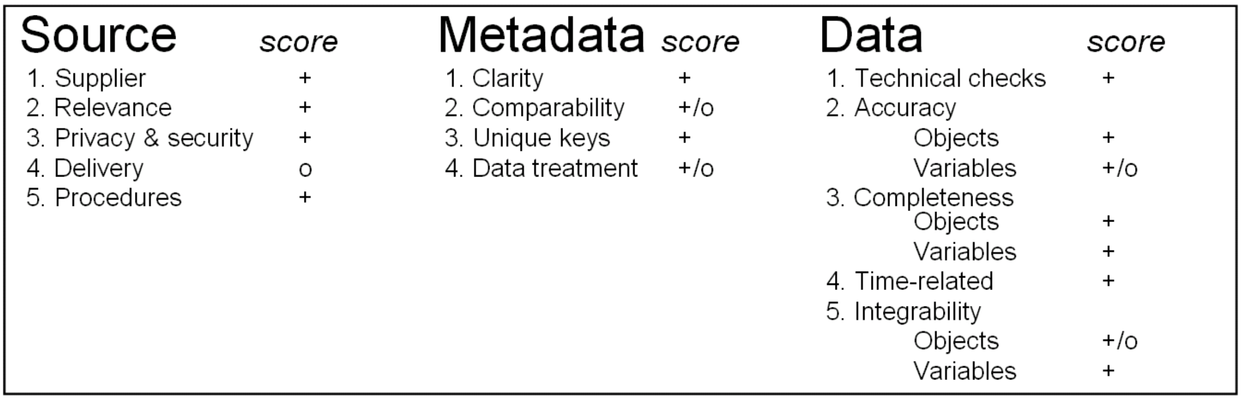
IV) The measurement methods output module. This module should be able to display, print and save measurement results.

More information on this topic can be found in Annex C and Annex D.

4.2 Quality Report Card

The previous chapter discussed -in detail- the quality indicators and measurement methods for determining input quality of an administrative data source. Although these detailed checks are needed to get a thorough impression of the (predicted) quality of a source, it is also useful to provide potential users with a quick global impression of the data sources quality. In this context the Quality Report Card (QRC) is proposed. Goal of the QRC is to present the outcomes of the indicators in an easy readable format at a dimensional level. Emphasis of the scores listed on the QRC could be DSQ or IOQ based. Detailed information, e.g. the individual scores for every method of each indicator, can always be additionally included in an underlying document. As such the QRC could be considered the top page of a quality document prepared for an administrative source. Since the QRC should give a global impression of the quality of a source, its metadata quality should also be included. For the example discussed in this paragraph, the evaluation results obtained with the Dutch Checklist (Daas et al., 2009) are shown. As a consequence of this choice, an example is shown in which quality is reported on three hyperdimensions: Source, Metadata, and Data. The first two are described in Daas et al. (2009), the latter in Deliverable 4.1 and in this document. In the view of the findings described for input quality in the beginning of this document (paragraph 2.1) optional reporting for Data on DSQ- or IOQ-level might be considered. The latter is ignored in the example shown in figure 1.

In figure 1 the hypothetical results for a data source are shown for the proposed card. By reporting only on quality at the dimensions level of each hyperdimension (5 for Source, 4 for Metadata and 5 for Data) a complete overview can be obtained at a glance. The dimensional scores in figure are the worst score obtained for an indicator in the dimension shown. The symbols for the scores used are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combining symbols with a slash (/) as a separator. For Source and Metadata this is logical because the indicators in these dimensions are nearly all quantitative ones (Daas et al., 2009). For Data also no numeric scores are given (figure 1) even though nearly all indicators are quantitative (see above). This is done deliberately because numeric scores tend to suggest more detail than is actually there. Furthermore how should a user interpret the difference between a score of 7.6 and a score of 7.7 and how should the indicators being part of a particular dimension be weighted against each other? For Data this implies that the values obtained for the indicators need to be mapped to the scores good (+), reasonable (o), and poor (-). Alternatives are good (☺), acceptable (😐), and unacceptable (☹). Whenever an unacceptable score is found, the user should obtain advice on the subsequent step(s) to take.



Source	score	Metadata	score	Data	score
1. Supplier	+	1. Clarity	+	1. Technical checks	+
2. Relevance	+	2. Comparability	+/o	2. Accuracy	
3. Privacy & security	+	3. Unique keys	+	Objects	+
4. Delivery	o	4. Data treatment	+/o	Variables	+/o
5. Procedures	+			3. Completeness	
				Objects	+
				Variables	+
				4. Time-related	+
				5. Integrability	
				Objects	+/o
				Variables	+

Figure 1. Draft version of a Quality Report Card for an administrative data source. Source and Metadata refer to the hyperdimensions proposed by Statistics Netherlands (Daas et al., 2009). The DSQ- or IOQ-option for Data is not included in this example.

5. FUTURE WORK

After constructing measurement methods for the indicators identified for the input quality of administrative data, the next steps in the research performed in WP4 are the following. First, the measurement methods that the NSI can perform need to be tested on several administrative sources used by the NSI's involved in WP4. This is needed to assure that the methods are correctly defined and provide valuable information. Methods that fail these tests need to be adjusted and in the worst case removed. In addition, work will start on the construction of a tool that enables automatic determination of the measurement methods proposed in this document. Only the measurement methods that can be automatically determined should be included in the tool. The NSI of Slovakia (INFOSTAT) will perform this work with assistance from Statistics Netherlands.

As stated in the previous deliverable (Deliverable 4.1), part of the research effort of WP4 focuses on the development of an overall approach to the quality assessment of administrative data sources (Laitila et al., 2011). This work is very valuable (see chapter 2) and will continue in the remainder of WP4. It is important that the new insights into the scope of input quality are included in the development of an overall approach to the quality assessment of administrative data sources and the final deliverable of WP4.

REFERENCES

Arts, C.H., Bakker, B.F.M., van Lith F.J. (2000) Linking administrative registers and household surveys, *Netherlands Official Statistics*, 15, 16-21.

Bakker, B. (2010) Micro-integration: State of the Art. Paper for the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, USA, p282.

Daas, P.J.H., Ossen, S.J.L. (2011) Metadata Quality Evaluation of Secondary Data Sources. *International Journal for Quality Research,* 5 (2), 57-66.

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, Statistics Finland and Eurostat, Helsinki, Finland.

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.

Deliverable 4.1 *List of quality groups and indicators identified for administrative data sources*, First deliverable of Workpackage 4 of the BLUE-ETS project.

ESSnet (2010) *List of quality indicators*. Deliverable 6.2 of Workpackage 6 of the ESSnet on Admin Data. http://essnet.admindata.eu/Document/GetFile?objectId=5272.

Eurostat (2005) Standard quality indicators. Quality in statistics, Seventh meeting, Luxembourg, May 23-24, Luxembourg.

Griffioen, A.R., van Delden, A., de Wolf, P-P. (2011) Key elements of quality frameworks, to be applied to statistical processes at NSI's. Paper for the 2nd European Establishment Statistics Workshop, 12-14 September, Neuchâtel, Switzerland

Laitila, T., Wallgren, A., Wallgren, B. (2011) Quality Assessment of Administrative Data. Paper for the 2011 European NTTS conference, Brussels, Belgium.

Olson, J.E. (2003) *Data Quality: the Accuracy Dimension*. Morgen Kaufmann, San Fransico, USA.

Schouten, B., Cobben, F., Bethlehem, J. (2009) Indicators of Representativeness of Survey Nonresponse. *Survey Methodology* **35**, 101-113.

Schulte Nordholt, E. Ossen, S.J.L., Daas, P.J.H. (2011) Research on the quality of registers to make data decisions in the Dutch Virtual Census. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Tague, N.R. (2004) Seven Basic Quality Tools. In: *The Quality Toolbox*. American Society for Quality, Milwaukee, Wisconsin, USA., p. 15. (link: http://asq.org/learn-about-quality/seven-basic-quality-tools/overview/overview.html).

Templ, M., Alfons, A. (2009) An application of VIM, the R package for visualization of missing values, to EU-SILC data. Forschungsbericht CS-2009-2, Vienna University of Technology, Austria.

Tennekes, M., de Jonge, E., Daas, P.J.H. (2011) Visual Profiling of Large Statistical Datasets. Paper for the 2011 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

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

Unwin, A., Theus, M., Hoffman, H. (2006) *Graphics of Large Datasets: Visualizing a Million*. Springer, New York, USA.

Zhang, L-C. (2009) A unit-error theory for register-based household statistics. Discussion Papers No. 598, December, Statistics Norway.

Annex A: Updated list of quality indicators for administrative data used as input

Dimension Indicators	Description	Examples
1. Technical checks	*Technical usability of the file and data in the file*	
1.1 Readability	Accessability of the file and data in the file	File is of an unknown format, is corrupted, contains an unfamiliar character set, or can not be decoded
1.2 File declaration compliance	Compliance of the data in the file to the metadata agreements	Metadata description not included or not available at the NSI, lay-out of file does not comply to lay-out agreed upon
1.3 Convertability	Conversion of the file to the NSI-standard format	File errors while decoding, corrupted data in file after conversion
2. Accuracy	*The extent to which data are correct, reliable, and certified*	
Objects		
2.1 Authenticity	Legitimacy of objects	Objects with invalid (syntactically incorrect) identification keys and objects with (syntactically correct but) wrongly assigned identification keys
2.2 Inconsistent objects	Extent of erroneous objects in source	Extent to which the objects in the source are (or can be made) internally consistent; especially important when the objects need to be converted (combined or split) by the NSI
2.3 Dubious objects	Presence of untrustworthy objects	Records of objects that can not with certainly be identified as objects belonging to the NSI population
Variables		
2.4 Measurement error	Deviation of actual data value from ideal error-free measurements	Errors resulting from invalid data collection by the administrative data holder including errors occurring during reporting, registration, and processing of the data.
2.5 Inconsistent values	Extent of inconsistent of combinations of variable values	Records with a combination of values for variables that is clearly erroneous
2.6 Dubious values	Presence of implausible values or combinations of values for variables	Records with values for variables or combinations of variables that are inconsistent and of which -at least- one must be erroneous
3. Completeness	*Degree to which a data source includes data describing the corresponding set of real-world objects and variables*	
Objects		
3.1 Undercoverage	Absence of target objects (missing objects) in the Source (or business register)	Objects active (in the reference period) but absent in source
3.2 Overcoverage	Presence of non-target objects in the source (or) In business register)	Source contains data for objects that do not belong to the target population (in the reference period)
3.3 Selectivity	Statistical coverage and representativity of objects	Incomplete coverage of target population in source, source only contains information for a very selective part of the population (e.g. only large retail companies in the south of the country)
3.4 Redundancy	Presence of multiple registrations of objects	Source includes multiple registrations of the same object (with exactly the same variable values)

Variables		
3.5 Missing values	Absence of values for (key) variables	Missing values for key variables, records without any values for variables
3.6 Imputed values	Presence of values resulting from imputation actions by data source holder	Data source holder imputes values without informing NSI and marking them
4. Time-related dimension	*Indicators that are time and/or stability related*	
4.1 Timeliness	Lapse of time between the end of the reference period and the moment of receipt of the data source	Data in source describes a period way in the past (e.g. 2 years ago), data set to old
4.2 Punctuality	Possible time lag between the actual delivery date of the source and the date it should have been delivered	Data source is delivered after the arranged date
4.3 Overall time lag	Overall time difference between the end of the reference period in the source and the moment the NSI has concluded that it can definitely be used	Data evaluation takes up a considerable amount of time which considerably delays the rapid use of the data
4.4 Delay	Extent of delays in registration the	Data reported to data source holder is not immediately updated in source. Changes in population and values registered are reported to the data source holder by the object concerned after considerable delay
Objects		
4.5 Dynamics of objects	Changes in the population of objects (new and dead objects) over time	Objects no longer belonging to the population are not removed, new objects are only added after multiple registration periods
Variables		
4.6 Stability of variables	Changes of variables or values over time	Variable composition changes between deliveries or values of reasonable stable (categorical) variables (such as NACE-code) changes back and forth between deliverables
5. Integrability	*Extent to which the data source is capable of undergoing integration or of being integrated.*	
Objects		
5.1 Comparability of objects	Similarity of objects in source -at the proper level of detail- with the objects used by NSI	Objects in source differ from those needed by the NSI and splitting up or converting them is very difficult
5.2 Alignment of objects	Linking-ability (align-ability) of objects in source with those of NSI	Degree of matching of objects in source to business register (or other base registers) of NSI, number of mismatches
Variables		
5.3 Linking variable	Usefulness of linking variables (keys) in source	Linking variables of objects in data source differ from those used by NSI (foreign keys used), no key variables available
5.4 Comparability of variables	Proximity (closeness) of variables	Comparability of (total) values for key variables in the source and the values of similar variables in other data sources (registers and surveys) used by NSI

Annex B: Motivations for adjusting some of the quality indicators reported in Deliverable 4.1

Technical checks dimension

The indicator 1.4 ‘Data inspection results’ (Deliverable 4.1) has been dropped from this dimension. The most important reason for doing this is to provide a clear-cut focus on the technical aspects in this dimension. Removal also makes it easier to decide for a go/no go decision; the outcome of the ‘Data inspection results’ indicator is certainly not expected to be that clear (Tennekes et al., 2011). Another reason is that the methods for indicator 1.4 can be ‘absorbed’ into one or more of the other dimensions, aimed at specific quality dimensions. The systematic approach that this project has been and still is developing is gradually creating a more conclusive overview of the indicators in each of the five dimensions.

Accuracy dimension

Compared to the previous deliverable some changes have occurred in this dimension.

- The indicators ‘Identifiability’ and ‘Authenticity’ have been combined into a single indicator called ‘Authenticity’. An object may be illegitimate if its identification key does not conform to the correct syntax (i.e. ‘Identifiability’), or if it does not belong to the *universe* of objects according to a reference source (i.e. ‘Authenticity’). An example of the latter is the occurrence of a person identification number that has never been issued by the office responsible for that. As such, the original indicators 2.1 and 2.1 (Deliverable 4.1) identified a very similar situation; the occurrence of illegitimate / out-of-scope objects.

- The indicator ‘Consistency’ (Deliverable 4.1) is renamed to ‘Inconsistent objects’ which makes it more inline with a new variable indicator ‘Inconsistent values’. Moreover, the term *inconsistent* refers to hard / fatal errors that one can be certain of, in contrast to the term *dubious* which refers to implausible data that may be erroneous.

The variable indicators ‘Reporting, Registration, and Processing errors’ have been combined into a single indicator called ‘Measurement error’. One reason is that, often though not always, one has to consult the data source holder regarding these indicators. Another reason is that, while it is useful for the data source holder to improve the quality if the different sources of measurement errors (as distinguished from each other), it is not always so for the NSI who uses the data for statistical purposes. The extent of the combined measurement errors may well be the only thing that matters from the NSI’s point of view.

Time-related dimension

A new indicator Delay is added. This indicator applies to both objects and variables since a delay in registration always affects the two at the same time. As a result of this the original description of the indicators Dynamics of objects and Stability of variables are revised to include this change accordingly.

Integrability dimension

No changes.

Annex C: Software implementation of the Quality Report Card Indicators for Administrative data

Discussion paper prepared by Alexander Láng, INFOSTAT

Introduction

This document has the ambition to answer to a list of questions for INFOSTAT mailed by Piet Daas on June 6, 2011 (“Wp4-questions for INFOSTAT” see attached below). The questions are related to the development of a software tool for Quality indicators listed for the brainstorming meeting of WP4 participants on “Measurement methods for the indicators of input quality” on June 16-17, 2011 in Stockholm. In the following text we will try to provide clear answers to the questions raised and clarify the position of INFOSTAT in WP4.

Software implementation

In the years 2006-2007 INFOSTAT has participated as a member of a consortium on the project of “Automated transmission of Administrative data to the environment of the Automated Statistical System (ASIS)” of the Statistical Office of Slovak Republic (SOSR).

The Administrative Data System (ADS) now provides for periodical transfer of administrative data from all major administrative sources in Slovakia – ministries, tax directorate, customs office, social insurance, geodetic and cartographic institute, etc.

Our main contribution in the ADS project was the enhancement of the ASIS Metainformation System (METIS) to include the description of external data sources and transformation rules for storing of external data in ASIS Oracle based source databases.

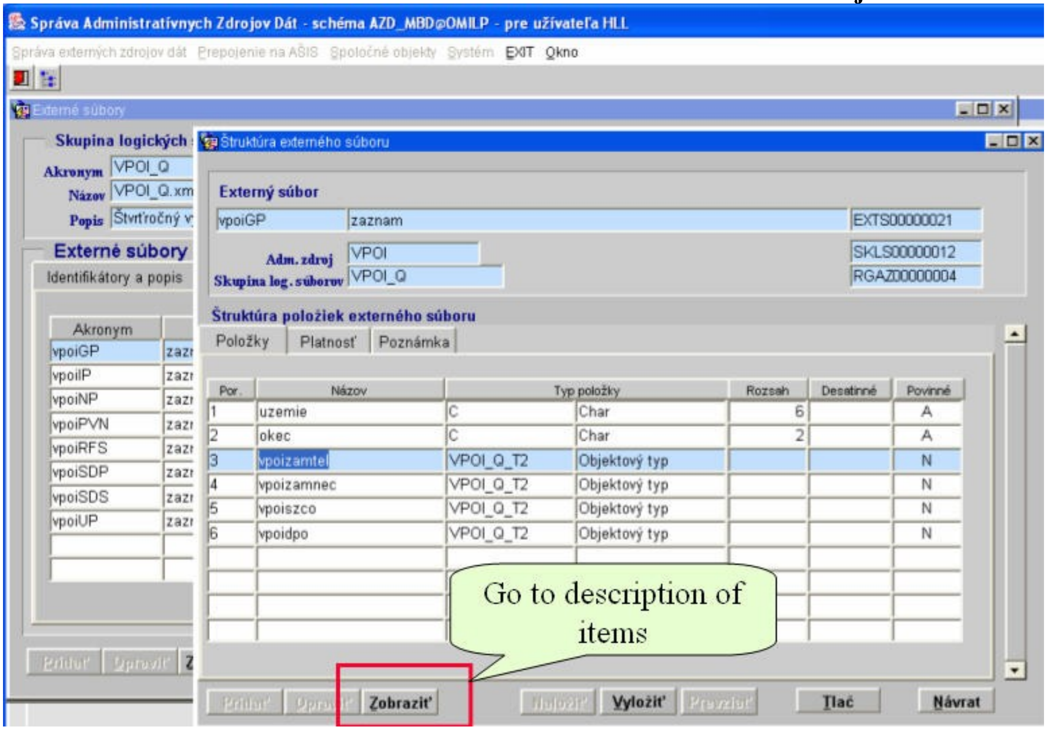
We also have solved problems with inputs from various communication channels (FTP, web interfaces, CD-ROMs, etc.) and different file formats (flat files, CSV files, dBase files, Excel spreadsheets, XML files). All communication of the ADS administrators takes place through a user-friendly Graphical User Interface (GUI).

In Fig. 1 we illustrate a description of administrative data from the Social insurance system, which is an XML file.

In Fig.2 we illustrate the description of Social insurance data from Fig. 1 as they will be stored in the ASIS source database.

We believe that the application software for the Quality Report Card (let’s label it as AQRC) should have a user-friendly and intuitive GUI where the user in successive steps invokes various modules to prepare and run quality measurements of administrative files.

Discussion topic: There is a variety of programming languages for developing GUIs but probably the most universally available package in NSIs is probably Microsoft Visual Basic.



Skupina logických

Akronym	VPOI_Q
Názov	VPOI_Q.xm
Popis	Štvrťročný v

Externé súbory

Identifikátory a popis

Akronym	
vpoiGP	zazr
vpoiIP	zazr
vpoiNP	zazr
vpoiPVN	zazr
vpoiRFS	zazr
vpoiSDP	zazr
vpoiSDS	zazr
vpoiUP	zazr

Externý súbor

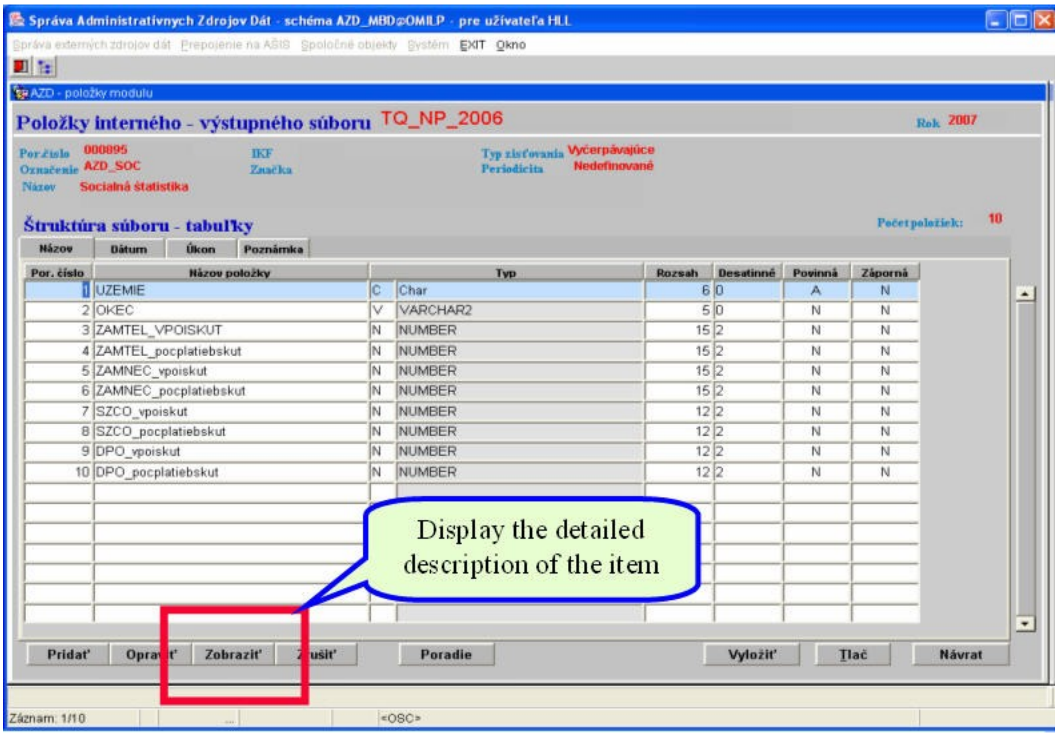
vpoiGP	zaznam	EXTS00000021
Adm. zdroj	VPOI	SKLS00000012
Skupina log. súborov	VPOI_Q	RGAZ00000004

Štruktúra položiek externého súboru

Por.	Názov	Typ položky		Rozsah	Desatinné	Povinné
1	uzemie	C	Char	6		A
2	okec	C	Char	2		A
3	vpoizamtel	VPOI_Q_T2	Objektový typ			N
4	vpoizamnec	VPOI_Q_T2	Objektový typ			N
5	vpoiszco	VPOI_Q_T2	Objektový typ			N
6	vpoidpo	VPOI_Q_T2	Objektový typ			N

Go to description of items

Fig. 1 – XML Source data description



Položky interného - výstupného súboru TQ_NP_2006

Rok 2007

Por.čislo	000895	IKF		Typ zisťovania	Vyčerpávajúce
Označenie	AZD_SOC	Značka		Periodicita	Nedefinované
Názov	Socialná štatistika				

Štruktúra súboru - tabuľky

Počet položiek: 10

Por. číslo	Názov položky	Typ		Rozsah	Desatinné	Povinná	Záporná
1	UZEMIE	C	Char	6	0	A	N
2	OKEC	V	VARCHAR2	5	0	N	N
3	ZAMTEL_VPOISKUT	N	NUMBER	15	2	N	N
4	ZAMTEL_pocplatiebskut	N	NUMBER	15	2	N	N
5	ZAMNEC_vpoiskut	N	NUMBER	15	2	N	N
6	ZAMNEC_pocplatiebskut	N	NUMBER	15	2	N	N
7	SZCO_vpoiskut	N	NUMBER	12	2	N	N
8	SZCO_pocplatiebskut	N	NUMBER	12	2	N	N
9	DPO_vpoiskut	N	NUMBER	12	2	N	N
10	DPO_pocplatiebskut	N	NUMBER	12	2	N	N

Display the detailed description of the item

Záznam: 1/10 ... <OSC>

Fig. 2 – Internal data description

AQRC modules

We propose that AQRC should consist of the following modules:

- Input data description module – administrative data, NSI survey data (some measurement methods may require data from two or more (?) sources for comparison)
  - File types

  Discussion topic: Should we restrict the possible file formats to a small set of most used formats like flat files, DIF, CSV, XML,…?
  - Variable types – Character, Numeric
  - Variable position, size, number of decimal places
  - Possibility of empty (NULL) values
  - Variable roles – identifier (primary key), classification items

- Input data transformation module – reads input data and creates workspace files for data quality measurement methods module
  - Selection of input file/files
  - Selection of variables from input file/files
  - Transformation of input to workspace files

  Discussion topic: Should we allow also conditional transfer of records (by defining filters for creating subsets of input files)?

- Data quality measurement methods module

Discussion topic: Statistics Netherland prefers to implement the measurement methods using R – the free open source statistical package. From other NSIs we can mention Italy, Austria and the Slovak Statistical Office where we know about statistical applications using R. INFOSTAT has no objections to this proposal.
  - Selection of a method
  - Selection of variables for the selected method

- Measurement methods output module – display, print or saving of measurement results.

AQRC measurement methods

In "Measurement methods for the indicators of input quality - brainstorm" document the methods are allocated to Quality indicators.

Our proposal is to consider more dimensions to these relations and structure them depending on types of variables:

- Single variable methods – identification keys, simple statistical calculations like average, median, standard deviation, coefficient of variance, …

- Multiple variables methods
  - Classification variable + numeric variables – e.g. distribution of turnover by NACE or size of the company
  - 2 classification variables and a numeric variable - e.g. table of turnover totals by NACE and company size groups
  - 2 numeric variables – ratios, statistical calculations (e.g. Coefficient of variation for total turnover per employee in the Eurostat Community Innovation Survey CIS 2008)

- same variables from more files – e.g. from administrative file and statistical file, or from the current file *t* and the previous period *t-1* file
- comparison of objects in a single file – check for duplicate records
- comparison of objects in two files – check for linking variables
- … etc.

Another aspect is that there may be special quality measurement methods for different kinds of data, e.g. financial data, demographic data, price indices, percentages, weights, etc.

Role of INFOSTAT in WP4 tasks

INFOSTAT has been developing application software for the Federal Statistical Office of Czechoslovakia since 1969 and for the National Statistical Office of Slovakia since 1993.
The core business and the main capacities of INFOSTAT are in the developing of statistical processing applications in ORACLE environment using PL/SQL, Oracle forms and Oracle Reports (e.g. complete Price statistics, Population Balance, Energy Balance, Intrastat survey, Innovation survey etc.). We also have experts developing applications in SAS (EU-SILC, part of Innovation survey, …), BLAISE (Labour Force Survey) and other programming languages – C#, Java, Visual Basic.

Concerning the R free open source statistical package we have some experience with using the R2.13.0Win version of the package. Therefore, should there be a consensus between consortium partners, INFOSTAT has no objection against using the R free open source package.

In short, INFOSTAT has the experience, capacity and skills needed for a successful completion of all tasks related to the development of AQRC for WP4 and is prepared to develop the application as described above and in compliance with the terms for deliverables 4.11 and 4.12 i.e. in M27 of the project.

References

Láng, A., Nogeová, Z., Hollý, P. (2008) Complex Application System for Data Transmission from Administrative Data Sources, PowerPoint Presentation, INFOSTAT, Bratislava, Slovakia.

Láng, A., Nogeová, Z., Hollý, P. (2008) Complex Application System for Data Transmission from Administrative Data Sources, Social Insurance Agency Case Study, PowerPoint Presentation, INFOSTAT, Bratislava, Slovakia.

Juriová, J. (2008) Quality Indicators for Selected Statistical Surveys (Monthly Construction Survey 2007, Community Innovation Surveys CIS 2004 and 2006), INFOSTAT, Bratislava, Slovakia

Juriová, J., Kľúčik M. (2009) Quality Indicators for Selected Statistical Surveys (2nd Stage), INFOSTAT, Bratislava, Slovakia

Templ, M., Filzmoser, P. (2008) Visualization of missing values using the R-Package VIM, Forschungsbericht CS-2008-1, Institut f. Statistik u. Wahrscheinlichkeitstheorie, Wien, Austria.

Templ, M., Alfons, A. (2009) An application of VIM, the R package for visualization of missing values, to EU-SILC data, Wien, Austria

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

Ehling, M. and Korner, T. (eds) (2007) Handbook on Data Quality Assessment Methods and Tools, Eurostat, Wiesbaden, Deutschland.

Annex D: WP4-questions for INFOSTAT

Introduction

On Thursday June 16 and Friday 17 the fourth WP4-meeting will be held in Stockholm. Since the start of the BLUE-ETS project and WP4 (in April 2010) considerable progress has been made. Most important result of WP4 is the creation is the list of quality indicators for the data in administrative data sources used as input in the statistical process. This result is described in Deliverable 1 of WP4 which has been accepted by the Project Officer of BLUE-ETS Ian Perry. An overview of the indicators identified is listed below:

Dimensions	*Quality indicators*	*Method description*
1. Technical checks	1.1 Readability	-Accessibility of the file and data in the file
	1.2 File declaration	-Compliance of the data to metadata agreements
	1.3 Convertibility	-Conversion of the file to the NSI-standard format
	1.4 Data inspection	-Results of preliminary data analysis
2. Accuracy	*Objects*	
	2.1 Identifiability	-Correctness of identification keys for objects
	2.2 Authenticity	-Correspondence of objects
	2.3 Consistency	-Overall consistency of objects in source
	2.4 Dubious objects	-Presence of untrustworthy objects
	Variables	
	2.5 Validity	-Correctness of measurement method used by DSH
	2.6 Reporting error	-Errors made by the data provider during reporting
	2.7 Registration error	-Errors made during data registration by DSH
	2.8 Processing error	-Errors made during data maintenance by DSH
	2.9 Dubious values	-Presence of inconsistent combinations of values
3. Completeness	*Objects*	
	3.1 Undercoverage	-Absence of target object in the source
	3.2 Overcoverage	-Presence of non-target objects in the source
	3.3 Selectivity	-Statistical coverage and representativiness of objects
	3.4 Redundancy	-Presence of multiple registrations of objects
	Variables	
	3.5 Missing values	-Absence of values for (key) variables
	3.6 Imputed values	-Values resulting from imputation actions by DSH
4. Time-related dimension		
	4.1 Timeliness	-Time between end of reference period and receipt of source
	4.2 Punctuality	-Time lag between the actual and agreed delivery date
	4.3 Overall time lag	-Overall time difference between end of reference period and the moment NSI concluded that the source can be used
	Objects	
	4.4 Object dynamics	-Usefulness of source to identify population changes
	Variables	
	4.5 Variable stability	-Consistency of variables or values over time in source
5. Integrability	*Objects*	
	5.1 Object comparability	-Similarity of objects in source with the NSI-objects
	5.2 Alignment	-Linking-ability of objects in source with NSI-objects
	Variables	
	5.3 Linking variable	-Usefulness of linking variables (keys) in source
	5.4 Variable comparability	-Proximity (closeness) of variables

Next stage

The next stage of the project is the creation of measurement methods for the quality indicators which, after acceptance by the WP4-members, should be implemented in a software tool/program. Statistics Netherlands has proposals for many of the methods for the quality indicators. We also found that some of the indicators can not be measured at all; but for these information possibly could be obtained from the provider of the data (the Data Source Holder). During the coming meeting in Stockholm these proposals will be presented and discussed with the other members of WP4 (Norway and Sweden; Italy unfortunately could not attend).

When consensus is reached on the measurement methods, these results will be written down in a document that is the basis for Deliverable 4.2 of WP4. This also initiates the next stage of WP4, the creation of a software tool/program for the indicators that can be measured. The latter tool will increase the chance that the results of WP4 are used and make it much easier for the users to measure the considerable number of indicators in a consistent way.

Software implementation

Before the methods can be implemented a considerable number of questions need to be answered, the most important ones are:

Questions	*Answers or considerations*
What should be implemented?	Only the measurement methods for data! Certainly for input!, but what about processing and output?
How to implement?	1) As a separate program (in what language?; e.g. C#, java) 2) As an 'add-on' to a statistical software program (SPSS, SAS, R etc.)
Who should create the tool?	Slovakia, Statistics Netherlands, others? (in what languages can they program?)
Who will use the tool?	European NSI's (but what software do they want?) Certainly Netherlands and Italy, Norway, Sweden…

Requirements

There are also some requirements to which the software certainly has to meet. The program should:

- be able to cope with large datasets in various formats
- not be expensive (available for free is to be preferred)
- preferably be used by as many NSIs as possible

Future decision

The preference of Statistics Netherlands is implementing the measurement methods as an add-on to R (the free open source statistical package). This is also used by statisticians in Norway, Italy, and Sweden. This need to be discussed and analysed within research activity which will be held throughout the second period of Blue-Ets project .