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 4.1

Dissemination level: PU

List of quality groups and indicators identified for administrative data sources

Authors:

Piet Daas, Saskia Ossen, Martijn Tennekes (Statistics Netherlands)

Li-Chun Zhang, Coen Hendriks, Kristin Foldal Haugen (Statistics Norway)

Antonio Bernardi, Fulvia Cerroni (Italian National Institute of Statistics)

Thomas Laitila, Anders Wallgren, Britt Wallgren (Statistics Sweden)

Final version

10 March 2011

EUROPEAN COMMISSION
European Research Area

SEVENTH FRAMEWORK PROGRAMME

Funded under Socio-economic Sciences & Humanities

Deliverable 4.1

List of quality groups and indicators identified for administrative data sources

Summary

This document contains a list of quality indicators identified for administrative data when used as an input source for the statistical process of National Statistical Institutes. The indicators measure the quality of the data in an administrative source and are grouped according to the following five general dimensions of quality: Technical checks, Accuracy, Completeness, Integrability, and Time-related dimensions. If applicable, a distinction has been made in each dimension between quality indicators specific for objects (such as units and events) and for variables. The list and grouping of the quality indicators forms the basis of the future work of workpackage 4 of the BLUE-ETS project, which is the development of a quality-indicator instrument for administrative data sources used in the statistical process.

Index

INTRODUCTION

Many National Statistical Institutes (NSI's) want to increase the use of administrative data (i.e. registers) for statistical purposes. To enable the use of administrative data sources by NSI's a number of prerequisites have to be met. These are in decreasing order of importance:

1. Availability of administrative data sources
2. Conformation of the NSI to a set of preconditions to enable the use of administrative data sources on a regular basis
3. Availability of methods to evaluate the statistical usability (i.e. quality) of administrative data sources in a standardized way.

To enable the use of administrative data for statistical purposes, relevant administrative data sources need to be available in the home country of the NSI. Because of the increase in the use of information and communication technology in public administrations (e-Government), this should not be a problem in most countries nowadays (Socitm, 2002). One may expect that at least some of these data sources are of potential interest for NSI's (Wallgren and Wallgren, 2007).

The second prerequisite is the topic of an excellent 'best practice paper' by the NSI's of the Nordic countries (Unece, 2007). This paper gives a thorough overview of the preconditions required to enable an NSI to extensively make use of administrative sources in statistics production. The preconditions are: 1) legal foundation for the use of administrative data source, 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. Conformance to these conditions will enable an NSI to use administrative data for statistics on a regular basis. No additional work is needed in this area.

When the two 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 the statistical usability (i.e. the quality) of these sources 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 administrative data 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 in a quick, straightforward, and standardised way. As yet, however, no standard instrument or procedure is available for such data sources (Berka et al., 2011; Daas et al., 2010; Frost 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 needs to be developed that is practical, robust, efficient, and applicable to a whole range of administrative data sources. To achieve this two important 'hurdles' have to be taken. These are:

1) the identification of the quality 'components' that determine the input quality of administrative data sources.
2) the development of an overall approach to the determination of the quality of administrative data sources.

The first 'hurdle' focuses on the identification of the quality components that make up the input quality of administrative data sources. It is a topic that has not received a lot of attention in statistics (Daas et al., 2010 and Annex A). It is essential that the components of input quality for administrative data are identified because it will enable NSI's to quickly decide -preferably immediately after receipt- if the administrative data source conforms to there needs. This is particularly important for NSI's that only recently have started to use administrative data source, to get a better grip on the dependency risk (Daas et al., 2009). The second 'hurdle', the development of an overall approach to the determination of the quality of administrative data sources, is also important and dependent on the first. However, next to the identification of the components of input quality, also a general applicable approach *needs* to be developed to asses the statistical usability of an administrative data source in the statistical process. The latter approach needs to cope with the effect of the use of the administrative data source on the production process and the quality of the final product (the output) of an NSI. With such a method one will be able to get a better grip on the advantages of the use of administrative data on the statistical process as a whole and the final output in particular.

The identification of the components of the input quality of administrative data sources, at the level of quality indicators, is the topic of this paper. The development of an overall approach is also currently studied but will be part of the next deliverables of WP4.

Quality determination

Administrative data sources can be used for many purposes by NSI's. For example, as input for a frame for sample surveys (e.g. the business register), as a source of auxiliary information, or as a replacement for data traditionally collected by a statistical survey (Unece, 2007; Wallgren and Wallgren, 2007). This tends to suggest that the quality of an administrative data source can only be established with relation to the intended use. For example, a source may be deemed of poor quality for providing data on the main variable under study, but the same source could -in another use- provide important auxiliary information. This poses a dilemma. On the one hand, statisticians want to known the quality of the source they are using as early as possible in the process (preferably prior to use) but on the other hand, the actual or intended use affects the way the quality of the source is perceived by the statistician. Thus, a general and useful system for quality assessment of administrative data sources can not originate from one specific application and the quality of statistics derived in that application (Laitila et al., 2011). Quality assessment of administrative data sources must therefore focus on i) information already available for the source and on ii) information that is the result of a systematic analysis of the source.

Looked upon in this way, it becomes natural to think about administrative data sources as inputs in a production system, i.e. inputs to a production function. Raw material can in general not be directly used in the production process; it has to be prepared, such as the cleaning of recycled fibre in paper mills. Substitute raw materials may imply a difference in the quality of the final product and the efficiency of the production process, e.g. virgin fibre gives stronger paper than recycled fibre. Also, the production technique may not be defined for some inputs. The simple analogue to the paper mill example illustrates that the quality of administrative data sources has to be looked upon from two different views, from the view of the consumer of statistics and from the view of the producer of statistics. The consumer view concerns the quality of the final product, or the '*output quality*'. This is the way quality has been traditionally looked upon (Eurostat, 2009). The producer view concerns two problems: i) '*input quality*' – the preparations of the input needed for use in the production process and, ii) '*production process quality*' – the gains in production efficiency of using the input (Laitila et al., 2011). For the development of a system of quality assessment of administrative data sources each of these three concepts must be divided into a set of components describing different aspects on the quality concepts.

The first deliverable of the research of WP4 focuses on the quality of the start of the process of the use of administrative data sources by NSI's: the *input quality* of administrative data sources. We want to identify the components that ultimately determine the input quality of administrative sources. This will enable NSI’s to determine the statistical usability of administrative data sources -at a general level- *prior* to use. This type of quality assessment is sometimes referred to as *ex-ante*, for it attempts to *forecast* the quality of the final result early on in the process. We will use input quality in the remainder of this paper.

The earlier mentioned paper mill analogue also nicely illustrates the way the input quality of an administrative data source should be looked upon, as an *input* source in the process and its *expected* effect on the quality of the end product. In other words, when a paper mill starts using recycled paper as an input source the quality management team has to seriously reconsider the applicability of the standard set of quality indicators commonly used (and developed) for the wood of trees. It is to be expected that a considerable number of the traditional used quality indicators for this type of input source can not or only partially be applied to recycled paper. Let’s illustrate this with a few examples. For trees it can be expected that -apart from the price- the species (type of wood), thickness of the bark, and the amount of harvestable wood are important input quality indicators. For recycled paper the amount of contaminants and the amount of printing ink are examples of indicators that are expected to be of considerable importance for this type of source. Apart from these very different indicators, one can also expect that some indicators, such as the fibre content of the source and the average length of the fibres in the source, can be applied to both input sources. The most important input indicators for each type of source will be indicators that are indicative for the quality of the final product (either positive or negative). These are indicators that certainly need to be determined for every new batch of input material. One has to realize that for a *new* input source it can not be known in advance which indicators turn out to be the most informative because it is not known how the new source affects the production process and the quality of the final product.

Focus of this paper

This paper focuses on the identification of indicators for the quality of administrative data sources when used as *input* sources in the statistical process. Chapter 2 start by carefully looking at the composition of quality and input quality in particular. Here, the necessity of investigating every quality component is reviewed. This is done to assure that the effort of the research of WP4 focuses on the essential constituents of the input quality of administrative data. The results of chapter 2 form the basis for the subsequent identification of quality indicators. The latter are introduced and discussed in chapter 3. The report finishes with a preview into the future research of WP4 and the first results of feed-back from users at NSI’s.

1. INPUT QUALITY OF ADMINISTRATIVE DATA SOURCES

An instrument capable of determining the input quality of administrative data has to be efficient. It should not cost to much time and effort to determine the quality of the input because this could, theoretically, be determined every time a new delivery of the source or part of the source is received (Daas et al., 2010). It is therefore vital that the instrument developed focuses on the essential components of input quality; the key quality constituents of an administrative data source. Because these constituents can not be known in advance (see above), WP4 started by carefully reviewing how researchers in statistics and other research areas perceive and determine the quality of the secondary data sources they use as input for their work. This enabled us to identify the components of quality that are *generally* considered the most important by the users of secondary data sources. The results of this study are discussed below. The reader is referred to Annex A for more details.

1.1 Metadata quality

When the quality of a data source is determined two quality *domains* always need to be considered. These are quality in the Metadata and in the Data domain (Batini and Scannapieco, 2006). Metadata quality is not often studied independently of its Data counterpart but this approach has been successfully applied at Statistics Netherlands (Daas et al., 2008; 2010). The latter institute has even developed a checklist for the determination of the quality components in the Metadata domain (Daas et al., 2009). Major advantage of this approach is that the Metadata quality components of a source can be determined independently of its content and, as a result, does not have to be checked every time the data in an source is studied (Daas et al., 2010). By using the checklist a total of 31 delivery and conceptual metadata related quality indicators are evaluated for a source. The checklist also aims to minimize the effort and time required for evaluation.

Because a general method is already available to determine the quality of the metadata of administrative data sources (Daas et al., 2009), the study of the input quality components of administrative sources in WP4 solely focuses on the remainder; the quality components in the Data domain. Be aware that this choice does not suggest that metadata indicators will be excluded from the quality instrument that WP4 intends to develop. It is merely not required to construct a list of metadata quality components for the input of administrative sources because this work has already been done. Consequence of the work described in this paper is that only the quality indicators for the input quality of administrative *data* need to be identified and grouped.

1.2 Data quality

After setting the focus of the research of WP4 on the input quality of administrative data, the next step is to identify the essential *general* constituents. This restriction is essential for two reasons. The first reason simply has to do with efficiency. By limiting the number of components studied, the number of quality indicators in the ‘to be produced list’ is reduced and, as a result, less time will be spend determining them. The second reason has to do with keeping focus on the components studied and reducing the uncontrollable growth of the indicator list. It is not uncommon that studies that plan to create a list of quality indicators end up with a huge list containing all kinds of supposed ‘quality indicators’; see Daas et al. (2010) and Frost (2010) for examples. It should be realized that in such lists not all of the ‘indicators’ included are true quality indicators. A quality indicator is indeed a component of quality that can be measured, but a quality indicator can be measured by one or more methods (Batini and Scannapieco, 2006). Care should therefore be taken that the list created not merely becomes a list of measurement methods! This is an important distinction (Annex A). By first focussing on the essential components of data quality on a level higher than that of a quality indicator, a level commonly referred to as a quality dimension (Batini and Scannapieco, 2006), a line of investigation is followed that -as much as possible- attempts to prevents the uncontrollable growth of the indicator list and the mixing up of indicators and measurement methods.

1.2.1 Essential dimensions of data quality

Quality indicators that measure similar (related) components of quality are usually grouped into so-called dimensions of quality. Each dimension focuses on a specific part of quality (Batini and Scannapieco, 2006). It is at the level of dimensions that many quality studies and quality frameworks can be compared (Batini et al., 2009; Annex A). This was the starting point for a literature study in which the essential dimensions of the quality of the secondary *data* used by scholars in a whole range of research areas were determined. Many of these scholars use administrative data as input for their research. This study was performed by Daas and Ossen and its results are included as an annex to this paper (Annex A). The overall conclusion of the literature study was that four dimensions of data quality are *generally* considered essential; they are studied by nearly all researchers. These dimensions are Accuracy, Completeness, Coherence, and a so-called Time-related dimension. The most remarkable outcome is that Completeness is identified as a distinct dimension; in statistics it is commonly considered an integral part of the Accuracy dimension (see Eurostat 2009). Apparently for many users of secondary data, this dimension of quality is considered as important as Accuracy. This perception is very likely the result of a shift in the focus of the Accuracy dimension. Traditionally, the latter dimension focuses on the accuracy of the estimate; the output of the statistical process. Looked upon from the input point of view, this is clearly no longer holds for the Accuracy dimension. The main focus of this input dimension is the identification of errors (see paragraph 2.2.3). In addition, the literature study also recommended the inclusion of a Technical checks 'dimension' (Annex A). Although it can be debated whether or not a collection of technical checks should be called a dimension, this naming will be used throughout the remainder of this document for consistency reasons. The importance of the findings described in this paragraph are discussed below.

1.2.2 Quality indicators in dimensions

For each of the essential dimensions of secondary data identified, quality indicators specific for the input of administrative data -when used for statistics- needed to be developed. To stimulate this work an exercise was performed during the first WP4-meeting (see Minutes WP4 2010a for details). Goal of this exercise was to position quality indicators already identified for administrative data in other studies into a matrix resulting from the combination of the four essential dimensions of data quality (e.g. Accuracy, Completeness, Time-related dimension, and Coherence) and the three steps of the statistical process (Input, Processing, and Output). The Technical checks dimension was deliberately ignored here because indicators for this dimension only scarcely occurred in the lists studied; only Daas et al. (2008; 2010) have mentioned these. Within each of the four dimensions an additional differentiation was made between indicators specific for units (a term later replaced by the more appropriate 'objects') and for variables. The matrix is shown in Figure 1. To aid the attendants, two additional rows were added to the matrix: one marked 'Other dimension' and one marked 'Metadata domain'; the latter is not shown in Figure 1. These rows were added to enable members to allocate indicators to a dimension or domain of quality that, according to his or her opinion, did not belong to one of the four dimensions proposed.

Four lists of indicators were evaluated during the exercise, these were: i) quality indicators proposed for the input by ISTAT (Bernardi et al., 2010; identified with the symbol 'I'), ii) quality indicators proposed for the Data hyperdimension of the Statistics Netherlands framework (Daas et al., 2010; identified with the symbol 'C'), iii) quality indicators for administrative data included in the draft list of the ESSnet on Admin Data (Frost, 2010; identified with the symbol 'E'), and iv) the standard quality indicators of Eurostat (2009; identified with the symbol 'E_') used for statistical output. The complete lists are a part of the WP4 minutes (2010a). Attendees were encouraged to allocate as many indicators as possible.

	INPUT	PROCESSING	OUTPUT
Accuracy			E_A1 E_A6 E_A7
Completeness Units Variables	C1.1 C5.1 C2.1 C3.1 E6 E7 E22 E23 C6.1 E2 E8	 C4.1 C4.2 C5.2 C6.2	C10.3 C5.3 C6.3 C10.1
Time related dimensions	I3.4 E15 E17 E19		E_T1 E_T2 E_T3
Coherence	I3.2 C7.2	C8.1 C8.2 E14	E_C1 E_C2 E_C4 E_CH1
Other dimension			E34

I = ISTAT; C = CBS; E = ESSnet; E_ = Eurostat

Figure 1. Overall results for the indicator allocation exercise performed during the WP4-meeting in Heerlen. The results for the Metadata domain column are not shown (more details in Minutes WP4 2010a).

The allocation results of all members were compared and combined. Indicators that were allocated to the same cell by half or the majority of the WP4-members were definitely assigned to that specific cell. For these indicators it was additionally checked if a differentiation between indicators specific for units (objects) or variables could be made. The overall results in figure 1 display the overall opinion of the WP4-members. A total of 39 indicators were allocated.
A considerable number of conclusions can be drawn from these results obtained (see Minutes WP4, 2010a). For WP4 the results of the Input column in figure 1 are the most interesting. The general conclusions that are relevant for the research on input quality of WP4 are:

- A total of 17 indicators were allocated to the input column of the matrix. Of these indicators 2 were proposed by ISTAT, 6 by CBS and 9 by the ESSnet.
- No indicators were allocated to the 'Other' dimension of the *input* column. Not even indicators mentioned in the Technical checks part of the CBS-list were allocated here. Apparently, no consensus was reached on the position of these indicators. Perhaps the 'uneasy' perception of Technical checks as part of a separate quality dimension plays a role here. The only indicator assigned to the 'Other' dimension was ESSnet indicator 34; an indicator that focuses on costs aspects. Because it is not included in the input column, the latter indicator is not relevant for WP4.
- The dimension to which the majority of the indicators were allocated (in general and also for the input column) was the Completeness dimension (19 in total, 11 for input).
- Only in the Completeness dimensions it was possible to differentiate between indicators specific for units (objects) and variables. For some of the indicators this distinction could not be made.
- Of the four essential dimensions in the matrix, the Accuracy dimension was found to contain the least number of assigned indicators. In the Accuracy part of the input column (and processing column) no indicators were overall assigned. People did assign indicators to this dimension (and also to the input part), but there was no general consensus on the position of these specific indicators in the statistical process.

The exercise described above confirmed the importance of the four essential dimensions for input quality, no other dimensions -apart from the Technical checks (see Minutes WP4, 2010a,b and paragraph 2.2.4)- are apparently needed. The results also provided hints for potential useful quality indicators in the Completeness, Time-related, and Coherence dimensions. Moreover, the results also suggested a distinction between input and processing quality from an indicators point of view. The indicators specific to the linking of records were generally assigned to the processing column and *not* to the input column. This suggests that any steps prior to the linking of units (objects) can be considered as belonging the input part of the statistical process. The importance of this distinction will become even more clear in the next paragraph. What the exercise did *not* confirm was the requirement for a Technical checks dimension and examples of input indicators for the Accuracy dimension. The first topic is discussed in paragraph 2.2.4. The second topic will be discussed in the next paragraph. In it a line of research is described that enabled us to identifying potential quality indicators for the input in the Accuracy and Completeness dimensions.

1.2.3 Sources of error in administrative data sources

As is clear from figure 1, the most challenging dimension of input quality is Accuracy. No consensus existed on assigning indicators to the input part of the latter dimension. There are fortunately other sources of information available that can assist us in this quest. For statistical surveys it is a well known fact that the composition of the quality indicators in the Accuracy dimension is related to the sources of error occurring in the data collection process (Batini and Scannapieco, 2006; Bethlehem, 2009; Groves et al., 2004).

Based on the sources of error scheme of Groves et al. (2004), which was originally developed for statistical surveys, and realising that administrative data source contain their own unique sources of error (Wallgren and Wallgren, 2007), Bakker (2010) has constructed a sources of error scheme specific for administrative data sources. Expanding on this, Li-Chun Zhang (2010a,b) has created a two-stage version of this scheme. In the first stage, the errors occurring in the collection and processing of a single source of administrative data are identified (figure 2); this stage was named ‘single-source statistical micro-data’ because the data is collected and processed independent of other data sources. The scheme essentially lists the errors that *can* occur during the data collection and maintenance process at the administrative data holder and in the first part -more or less the checking phase- of the statistical process at an NSI. The scheme shown in figure 2 identifies the sources of error up to the point at which administrative data is linked to other (statistical) data. The scheme for the second stage discerns the sources of error that result from the integration of statistical micro-data; a step performed at an NSI. This scheme is not discussed here because it is not part of the input phase; it is however included in Zhang (2010b) and Minutes WP4 (2010a).

The single-source statistical micro-data scheme (figure 2) identifies potential sources of error related to measurement, at the level of variables, and to representation, at the level of objects (units). During WP4-meetings this scheme proved not only useful for identifying Accuracy related quality indicators, but also for indicators in the Completeness dimension (see Minutes WP4, 2010a,b for details). Many of the indicators included in the final list for the Accuracy and Completeness dimensions (see chapter 3) were derived from the scheme shown in figure 2.

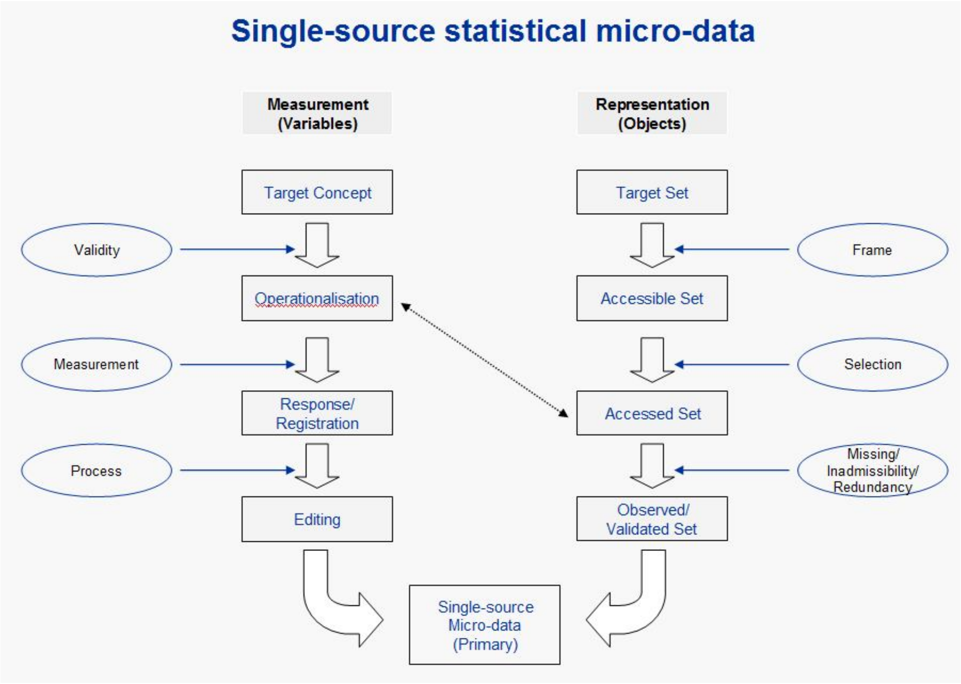


Figure 2. Overview of sources of error in single-source statistical micro-data. This is the first stage of a total of two. In the ellipses the possible sources of error in each step are identified (from: Zhang, 2010b).

1.2.4 Additional considerations

All of the results described above and the knowledge, experience, and ideas of the WP4-members provided enough information to start constructing a list of input quality indicators in the Accuracy, Completeness, Coherence, and Time-related dimensions of administrative data. In addition to the decision to include of a Technical checks dimension (see below and Minutes WP4, 2010c) several other changes were made. The most important considerations are described below.

Technical checks dimension

Since not much was known about the actual requirement for a Technical checks dimension (apart from its inclusion in the Data hyperdimension list of Statistics Netherlands and its suggestion in the literature study) it was decided to take a more specific look at the current steps performed during the evaluation of a new administrative data source at an NSI. As an example, the evaluation of a new administrative data source in Norway were carefully noted; the report of this study is part of Minutes WP4 (2010b). This exercise definitely confirmed the need for a Technical checks dimension. It was found that a considerable number of the evaluation steps performed at the start of the evaluation were identified as Technical checks (see Minutes WP4, 2010b). Readability (accessibility) of the file, compliance of the data to the metadata included, and correct conversion of the file to the NSI internal format are a few examples of the technical checks found. From this practical example, it was obvious that Technical checks (and these indicators) unquestionably needed to be included in the list of essential input quality components. To remain consistent with the naming of the other dimensions, a Technical checks dimension was included.

Coherence vs. Integrability

Upon careful consideration, the Coherence dimension was found to be composed of two parts. The first is the ‘within coherence’, containing indicators that focus on the (internal) consistency of the variables in the source; e.g. comparison of the variables at the record level. The second part of Coherence is the ‘between coherence’ which is the comparability between the data in different sources or in separately delivered parts of a source (see Minutes WP4, 2010b). Since the group found that the ‘within coherence’ was very much related to Accuracy at the input level, it was decided to include the indicator(s) belonging to this part of the Coherence dimension into the Accuracy dimension. The remainder were kept in the Coherence dimension, which was now solely composed of ‘between coherence’ indicators. Because of this it was decided to rename the Coherence dimension to Integrability. The new name for this dimension more clearly stated what the remainder of the coherence dimension was about, viz. “how well can the administrative data source be integrated into the statistical process/system of the NSI?”.

Time-related dimension

The time-related dimension naturally contains the well-known indicators timeliness and punctuality (Eurostat, 2009) which are, in the input phase, applicable to a specific delivery of administrative data. However, when a data source is regularly used by an NSI, stability of the data in the source over time also becomes an important topic (Minutes WP4, 2010a). The latter is sometimes also referred to as comparability over time (Minutes WP4, 2010b). Because of this, indicators related to stability (or changes) of objects and variables over time needed to be included in the Time-related dimension. For objects, an indicator related to population changes (population dynamics) covered by the source needs to be included. For variables, another type of indicator was identified. Over time, the values of particular variables (such as turnover) will, of course, change between subsequent deliveries. It is, however, important that the variable composition in the source remains stable and that the values of some of the variables (such as the NACE code) do not change back and forth between subsequent deliveries. Typical for stability indicators is that the data in the delivery under study is compared to several of the previous deliveries.

2. LIST OF INDICATORS

Based on the information provided above and additional discussions between the WP4-members (partly described in Minutes WP4, 2010b), a list of quality indicators was constructed that is specific for administrative data used as input for NSI's. The list is shown in Table 1 (pages 13-14). For the future research of WP4 it is important that the list contained *all possible* identifiable quality *indicators* for administrative input *within* the set of selected dimensions. The Latin proverb 'Melius abundare quam deficere' applies here (in English: 'Better too much than not enough'). This seems to contradict the restrictive line of reasoning followed before (see paragraph 2.2) but it is not. Returning to the analogy of the paper mill, when a new input source is going to be used in the production process -such as recycled paper- it is essential that the whole range of possible *indicators* that *could be* applied to this type of source is identified. The latter should of course be done *within* the context of the commonly used dimensions of input quality. As long as the list is solely composed of possible indicators (and not measurement methods), a bit more is better than too little. This is also important for the future research of WP4, as it can -at this point in time- not be known in advance which input indicators will be most indicative to the quality of the output; either in a positive or in a negative way. By not limiting the number of potential *indicators* in each dimension, it becomes more certain that the important input indicators are included in the list.

For all dimensions in table 1, with the exception of the Technical checks dimension, a differentiation is made between indicators specific for objects and for variables. In the table also a definition of each dimension is provided. The table also includes a description for each indicator and one or more examples to illustrate its application range. Measurement methods are not considered part of this deliverable (see chapter 4).

Technical checks

A total of 4 indicators are included in the Technical checks dimension. It 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, this dimension also contains an indicator that checks if the specific data delivery complies to its metadata-definition. The metadata can be included in the delivery, either as a separate file or as a header in the file (describing its content), but could also provided to the NSI in a separate process. In the Technical checks dimension also an indicator is included that expresses the results of preliminary data analysis (see Minutes WP4, 2010a). First results of an approach particularly suited for this task are included in Tennekes et al. (2011).

Accuracy

The indicators in this dimension all originate from the sources of error scheme in figure 2. This scheme identifies the sources of error when administrative data is used as input by 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 objects (units and events) in the source, while the variable indicators focus on the validity of the values provided. A total of 9 indicators are included; 4 for objects and 5 for variables.

Completeness

The indicators for objects in this dimension predominantly focus on coverage issues. The indicators for variables are related to missing and imputed values. Of the total of 6 indicators, 4 are object specific and 2 are indicators for variables.

Time-related dimension

The quality indicators in this dimension are all related to time. The timeliness and punctuality indicators apply to the delivery of the *individual* data file. In addition an indicator is included for the overall time lag of the delivery. This indicator measures the time lag between the reference period covered and the moment at which it can be used by the NSI. It therefore also includes the time required for evaluation.

The remainder of the indicators in the Time-related dimension are all stability related. The indicator for objects focuses on the dynamics of the population of objects in the individual file compared to

Table 1. 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
1.4 Data inspection results	Results of preliminary data analysis	Data profiling results, results of visual inspections, inconsistencies between multiple files delivered, value representation inconsistencies
2. Accuracy	*1) Closeness of the objects and variables to the exact/true objects and values defined, 2) The extent to which data are correct, reliable, and certified*	
Objects		
2.1 Identifiability	Correctness of identification keys for objects	Objects with invalid (syntactically incorrect) identification keys
2.2 Authenticity	Correspondence of objects	Objects with (syntactically correct but) wrongly assigned identification keys
2.3 Consistency	Overall consistency of 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
2.4 Dubious objects	Presence of untrustworthy objects	Records of objects that can not with certainly be identified as objects belonging to the NSI population
Variables		
2.5 Validity	Correctness of measurement method used by the administrative data holder for variable(s)	Errors resulting from invalid data collection by the administrative data holder
2.6 Reporting error	Errors made by the data provider during reporting	The data provider provides a wrong value for a variable (e.g. wrong start date of the business, wrong number of employees, wrong name of the company)
2.7 Registration error	Errors made during data registration by the administrative data holder	Wrong value due to mistakes in the registration process (e.g. misplaced comma, wrong spelling of an otherwise correct address)
2.8 Processing error	Errors made during data maintenance by the Administrative data holder	Value in a field is erroneously adjusted by the administrative data holder during data maintenance (data checks)
2.9 Dubious values	Presence of inconsistent combinations of values for variables	Records with values for 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 in the business register)	Objects active (in the reference period) but absent in source (or business register)

3.2 Overcoverage	Presence of non-target objects in the source (or in the business register)	Source (or business register) 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 administrative data holder	Administrative data holder imputes values without informing NSI and identifying 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
Objects		
4.4 Dynamics of objects	Usefulness of source to identify changes in the population of objects (new and dead objects)	Objects no longer belonging to the population are not removed, new objects are only added after multiple registration periods
Variables		
4.5 Stability of variables	Consistency of variables or values over time	Variable composition changes between deliveries or values of reasonable stable 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	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

those in previous deliveries. This can be either good or bad. The indicator for variables has a similar intention. Although the values of variables should of course change between subsequent deliveries, it is important that the variable composition covered by a source remains stable and that the values of some of its variables (such as the NACE code of a company) do not change back and forth between subsequent deliveries. In both cases the stability indicators focus on the changes of the data in an *individual* delivery compared to those in previous deliveries. Any other time related indicators that could be conceived of were, in essence, metadata related and as such already covered in the Metadata domain checklist; see Daas et al. (2009) for more details.

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 easy of linking the objects in the source to those commonly used by the NSI. The variable indicators either focus on the quality of the linking variable used or compare the closeness of the values in the source to the values of similar variables in other sources. A total of 4 indicators are included in this dimension, 2 for objects and 2 for variables.

3. FUTURE WORK

After identifying the indicators for the input quality of administrative data, the next steps in the research performed in WP4 go into 2 direction. One direction continues the work on input quality while the other part focus on the development of an overall approach to the quality assessment of administrative data sources (Laitila et al., 2011).

The work on input quality now go into a more practical direction. The first results of internal reviews by users of administrative data sources at some of the NSI's involved in WP4 reveal that all indicators identified for input quality are considered important. These and other topics that will be studied now and in the near future in relation to input quality are:

i) The first thing that needs to be done is the development (and testing) of measurement or estimation methods for the quality indicators included in Table 1. It is absolutely essential that for each indicator at least one method is available to measure or estimate it. For some indicators such methods are already developed by others (e.g. indicators 3.2 Overcoverage and 4.1 Timeliness), for some additional information has to be provided by the administrative data holder (such as indicator 2.7 Registration error), and for others new methods have to be developed (such as the stability indicators 4.3 and 4.4).

ii) Also, the set of indicators and measurement methods should be reviewed for their validity by experienced users of administrative data sources at NSI's. The valid indicators (and methods) need to be tested on a number of sources used -for various purposes- by NSI's. An example of the latter is an administrative data source that is not only used to update the business register but also used for Structural Business Statistics. This will not only reveal the general applicability of the indicators in the list proposed (and the differences within and between countries), but will also provide clues on the importance of the indicators in relation to specific purposes and the sequence in which the indicators are evaluated. In other words, this could provide an answer to the questions: "are different sets of quality indicators needed when an administrative data source is used for different purposes?" and "are different approaches needed in different countries?". The answers to these questions are important in the development of an overall approach (methodology) for the determination of the quality of the input of administrative data; the second deliverable of WP4. Another subject that must be included in this line of research is the requirement of indicators for metadata in the evaluation. This could also vary between countries. Many of the issues raised in this topic have already been touched upon by WP4-members during the meetings (see Minutes WP4, 2010a,b).

iii) The results of the research of topic ii) will also provide clues to the importance of the indicators in the list. Some indicators might be applicable in almost every country for almost every source, while others might hardly ever be used or might found to be not very informative. While doing this also a first link can be made between the quality of the input and the output. In some aspects, this topic very much resembles the study of non-sampling errors in sample surveys, which are quite difficult to measure (Statistics Canada, 2010). Any research results of the topics i) and ii) in WP4 that could shed light on this very interesting research subject is a valuable contribution to statistical research.

iv) Finally, tools need to be developed to assist users. This is future research but, with this knowledge in mind, it is important to consider which of the steps in the evaluation of the input quality of administrative data sources could possible benefit from scripts or a more advanced software tool. This information should also be gathered during the work on topics i) and ii).

Regarding the work on the development of an overall approach to the quality assessment of administrative data sources, it is clear that this work should make use of all results obtained in this research area as much as possible (Berka et al., 2011; Daas et al., 2009, 2010; Frost et al., 2010; Laitila et al., 2011). The work described in this paper will certainly be an important contribution to the input part of the process reviewed.

The results of all the work described above will be included in the next deliverables of WP4.

REFERENCES

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.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A. (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys 41(3), Article 16, July.

Batini, C., Scannapieco, M. (2006) Data Quality: Concepts, Methodologies and Techniques. Springer, Berlin, Germany.

Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H., Schwerer, E. (2011) Quality Measures for Administrative Registers: Preliminary Results from the Austrian Census 2011. Paper for the 2011 European NTTS conference, Brussels, Belgium.

Bernardi, A., Cerroni, F., Di Giorgi, V. (2010) Analysis on Economic Fiscal Data for Statistical Uses. Paper for the Seminar Using Administrative Data in the Production of Business Statistics, Rome, Italy.

Bethlehem. J.G. (2009), Applied Survey Methods, A Statistical Perspective. John Wiley and Sons, Hoboken, USA.

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

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.

Eurostat (2009) ESS Handbook for Quality Reports. Eurostat Methodologies and Working papers, Office for Official Publications of the European Communities, Luxembourg, pp. 133-135.

Frost, J-M. (2010) WP6: Quality Indicators when using Administrative Data in Statistical Operations, 3rd Draft: Initial User Testing. ESSnet Use of Administrative and Accounts Data in Business Statistics, UK.

Frost, J-M., Green, S., Pereira, H., Rodrigues, S., Chumbau, A., Mendes, J. (2010) Development of quality indicators for business statistics involving administrative data. Paper for the Q2010 European Conference on Quality in Official Statistics, Statistics Finland and Eurostat, Helsinki, Finland.

Groves, R.M., Fowler jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (2004) Survey Methodology, Wiley Interscience, New York, USA.

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

Minutes WP4 (2010a) Minutes of the first Workpackage 4 meeting in Heerlen, Statistics Netherlands, dated 15 September.

Minutes WP4 (2010b) Minutes of the second Workpackage 4 meeting in Kongsvinger, Statistics Netherlands, dated 28 October.

Socitm (2002) Local e-government now: a worldwide view. Joint report of the Society of Information Technology Management and the Improvement & Development Agency, September.

Statistics Canada (2010) Non-sampling error webpage of Statistics Canada, 5 October version (http://www.statcan.gc.ca/edu/power-pouvoir/ch6/nse-endae/5214806-eng.htm).

Tennekes, M., de Jonge, E., Daas, P. (2011) Visual Profiling of Large Statistical Datasets. Paper for the 2011 European NTTS 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..

Wallgren, A., Wallgren, B. (2007) Register-based Statistics: Administrative Data for Statistical Purposes, Wiley, Chichester, U.S.A.

Zhang, L-C. (2010a) Assessment of uncertainty in register-based small area means of a binary variable. Presentation at the Workshop on Measurement Errors in Administrative Data, Mannheim, Germany, 14-15 June.

Zhang, L-C. (2010b) Developing statistical theories for register-based statistics. Qvintensen, Nr. 4, pp. 20 - 22, Swedish Statistical Association.

Annex A: In Search of the Composition of Data Quality in Statistics and Other Research Areas (Piet J.H. Daas and Saskia J.L. Ossen, 2010)

Deliverable 4.1: Annex A

In Search of the Composition of Data Quality in Statistics and Other Research Areas

Authors: Piet J.H. Daas and Saskia J.L. Ossen

Final version

10 March 2011

Deliverable 4.1: Annex A

In Search of the Composition of Data Quality in Statistics and Other Research Areas

Summary

National Statistical Institutes (NSI's) are increasingly using data collected by others for producing statistics. This has the disadvantage that the collection and maintenance of the data used is beyond the control of the NSI. It is therefore of vital importance that researchers try to determine the quality of the sources they are using. In determining the quality of a source two quality domains need to be distinguished: the Metadata and the Data domain. To assess the quality related to the Metadata domain Statistics Netherlands developed and thoroughly tested a checklist. Current research concentrates on the evaluation of the quality in the Data domain. In this paper an inventory is made of how scholars in different research areas deal with the determination of data quality. The overall conclusion of this inventory is that five input data quality characteristics dominate the study of Data quality in almost all research areas. These characteristics are: coherence, completeness, correctness, selectivity, and timeliness. These correspond to the data quality dimensions Coherence, Completeness, Accuracy, and Time-related dimensions.

Index for Annex A

Deliverable 4.1: Annex A

In Search of the Composition of Data Quality in Statistics and Other Research Areas

INTRODUCTION

National Statistical Institutes (NSI's) collect data for the production of statistics. Apart from the data obtained through surveys, NSI's are increasingly making use of data that is collected and maintained by other organisations for non-statistical purposes. An example of such a secondary data source is administrative data (Wallgren and Wallgren, 2007). Although this data is produced as a result of administrative processes within organizations, it is -very often- an interesting data source for NSI's. A fact more and more NSI's have realized during the last decade (Unece, 2007). Taking the lead in this development are the NSI's in the Nordic countries. In these countries secondary data is already the main data source for the production of official statistics (Statistics Finland, 2004; Unece, 2007; Wallgren and Wallgren, 2007).

A major advantage of using secondary data is the fact that it drastically reduces the costs of data collection and the response burden on enterprises and persons. It also provides large amounts of data and saves time that would otherwise be spent in collecting data (ESC, 2007). These advantages are not only seen by statisticians at NSI's but also by scientists in other research areas. Quite some researchers, such as medical scientists, have seen the advantage of re-analysing data that was collected by others (Skeet, 1991; Sørensen et al., 1996).

The use of secondary data, however, has some disadvantages as well. The most important one is the fact that the collection and maintenance of the data is beyond the control of the researcher. It is the data source keeper that manages these aspects. The same is true for the units and variables a secondary data source contains. These are defined by the administrative rules of the data source keeper and may therefore not be identical to those required by the researcher (Wallgren and Wallgren, 2007). Since the production of high quality output by using secondary data sources largely depends on the quality of the data in the source, it is of vital importance that researchers try to determine the quality of the input, i.e. the usability of the source for their particular purpose, prior to use and in an efficient way.

At Statistics Netherlands two domains are distinguished in determining the quality of a source: the Metadata and the Data domain (Daas and van Nederpelt, 2010). For the Metadata domain recently a quality checklist has been developed that enables a systematic and standardized assessment of the quality aspects belonging to this domain. By applying this checklist to several registers it has been shown that the checklist is a useful tool for identifying metadata quality related problems in registers (Daas et al., 2009).

Current research at Statistics Netherlands aims at developing a quality framework in which, apart from the Metadata checklist, also standardized methods are included for evaluating the quality aspects in the Data domain (Daas et al., 2010). This research is performed as part of the European BLUE-Enterprise and Trade Statistics (BLUE-ETS) project. This work started by identifying all characteristics relevant for the quality of secondary data sources when used as statistical input (Daas and Van Nederpelt, 2010). In this study the Object Oriented Quality Management (OQM) model was used to identify all potentially important *characteristics* of the quality of secondary data (Van Nederpelt, 2009). This resulted in the list of shown in Table 1. In this table explanatory text or questions are added to illustrate the interpretation of the characteristics more clearly.

Table 1. Characteristics identified for the data domain of quality for a secondary data source.

Characteristic	*Explanatory text*
Authenticity	Does the administrative unit refer to the intended (real world) unit?
Coherence	Coherence between the data of the items (per unit)
Completeness	Is all data delivered and can all data be accessed in the dataset? Including covering of units and completeness of the information for the items
Confidentiality	Confidentiality of the unit and item information in the dataset
Correctness	Correctness of the identification key used for the units and absence of measurement and other errors for the items
Detailedness	Level of detail for the item information in the dataset
Selectivity	Selectiveness of the coverage of the units and of the item information in the dataset
Stability	Stability of the content of the dataset over time, changes in number of records, changes in coverage and item information comparability over time
Timeliness	Recentness of the unite and item information in the dataset
Uniqueness	Presence and uniqueness of identification keys for the units and of the item information in the dataset (from an identification point of view)

Although table 1 contains a lot of characteristics of data that are of interest, the important questions are: i) should all these characteristics always be determined for a source? and ii) are any important characteristics missed? The first question is motivated by the consideration that the framework will only be effectively used in practice when data quality can be evaluated in a reasonable amount of time. The second question is the result of the fact that it is theoretical possible that the OQM-approach misses some important, not previously considered, characteristics.

In this paper we therefore look how statisticians and researchers in other fields deal with the determination of the quality of their (secondary) data sources. Unfortunately not every study uses the same set of ‘terms’ to express their findings. We therefore identify all *characteristics* of data quality observed in each study to enable a proper comparison of their findings. Stated differently, we do an inventory of the work done in several research fields in order to gain insight into the set of characteristics that should be incorporated in our data quality framework.

To reach this aim this paper is structured as follows. We start in chapter 2 by giving more insight into different interpretations of quality as we found that quality can be interpreted in many different ways. We more specifically discuss the different levels of detail at which quality can be considered. Since, in this paper we are interested in determining the quality of administrative data used as input for the statistical process, we also stress the difference between input, throughput, and output quality. Chapter 3 gives an overview of the input quality aspects viewed upon by NSI’s. In chapter 4 the aspects of input quality considered in other, non-statistical, research areas are discussed. In the final chapter, we conclude by using the obtained insights to identify which characteristics of input data quality are most commonly used in practice and link these characteristics to the more commonly used dimensions of data quality.

Please note that we have tried to make the literature inventory as complete as possible. Because of the broadness of this field of research and the limited time available, there is however always a chance that some publications were missed. Papers presenting the same findings in a different context are deliberately ignored.

1. QUALITY

1.1. What is quality?

Searching the internet for the answer to the question “What is quality?” leads to many results. These results reveal that quality is interpreted in many (slightly) different ways (Wade, 2005). The definition most commonly observed is the one of Joseph M Juran who defined quality as "*Fitness for use*" [1] (Juran, 2004). There are, however many (slightly different) definitions out there. From this it is clear that quality is a multifaceted or -more accurately- a multidimensional, concept (Batini and Scannapieco, 2006). The term “quality” is used in many different ways, for instance for products, for processes, and for services (Ehling and Körner, 2007). In this document we have limited the scope of our work to quality from a product point of view. The International Organization for Standardisation (ISO) has defined product quality in ISO 8402 – 1986. In here it is stated that quality is *“the totality of features and characteristics of a product….that bear on its ability to satisfy stated or implied needs*” (ISO, 1986). This definition of product quality is also used by Eurostat (2003a) for statistical products. Product should in this paper be interpreted as administrative data used as input to the statistical process of an NSI.

1.1.1. Product quality: different levels of detail

Our literature study revealed that product quality can be looked upon at different levels of detail. Figure 1 provides an overview of these levels. Knowledge about the different levels of detail is important to interpret the different aspects of quality mentioned in literature; different studies do not always refer to the same level of detail. Because of this difference we therefore decided to translate the findings in each study to the characteristics of quality (Van Nederpelt, 2009).

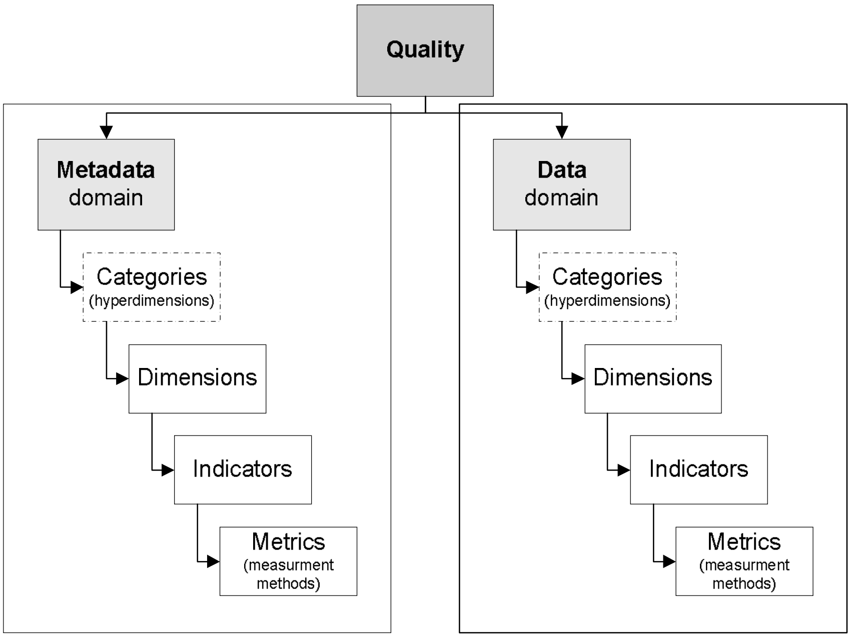

Quality
Metadata domain
Categories (hyperdimensions)
Dimensions
Indicators
Metrics (measurment methods)
Data domain
Categories (hyperdimensions)
Dimensions
Indicators
Metrics (measurment methods)

Figure 1. Overview of the composition of quality

At the lowest level of detail of product quality a differentiation needs to be made between Data and Metadata quality (Batini and Scannapieco, 2006). In this paper we will concentrate on the quality of

[1] Later in life, Joseph Juran changed this definition to the combination of: 1) The degree to which costumer needs are met and 2) the absence of error (Juran, 2004).

the Data domain as for the Metadata domain already a quality checklist has been developed at Statistics Netherlands (Daas et al., 2009). During the literature study we also found that most quality research tends to focus on the quality of the data. Metadata quality is hardly ever studied on its own. If it is studied it is usually included as part of a framework constructed for the determination of quality in general. As a result, the line between the Data and Metadata domain of quality is not always sharply drawn in many of the studies found. We nevertheless try to make the distinction between metadata and data quality aspects as we want to use the results of this paper for developing a standardized method for determining data quality.

At a more detailed level of quality, dimensions come into play. In all of the papers found, researchers have identified several dimensions of quality. Wang et al. (1995) performed an extensive study in which they particularly focussed on the dimensional composition of quality. For clarity, this concerned both the Data and Metadata domain of quality. The most important finding of Wang et al. (1995) is that no general agreement exists on the number and ‘types’ of dimensions discerned. These findings will be confirmed in the remainder of this paper. This again prompted us to translate the findings in each study to the characteristics of quality of Table 1. According to Van Nederpelt (2009) the characteristics of quality are located at level of detail very close to that of dimensions (Daas and Van Nederpelt, 2010).

The measurable part of quality, an even more detailed level, is called an indicator (Ehling and Körner, 2007). In general, each dimension of quality contains several indicators. Every indicator measures a specific -preferably different- aspect of the quality dimension concerned. Metrics, measures or estimates of quality, are needed to determine the value of a quality indicator. The measurement methods used can be either qualitative or quantitative. A quality indicator is measured (or estimated) by at least one method, but sometimes a combination of two or more methods is also used.

Occasionally, ‘intermediate’ detail levels of quality are introduced by researchers. This is particularly the case when a lot of dimensions are identified; see Wang and Strong (1996) and Daas et al. (2008) for examples. These intermediate levels are composed of grouped dimensions and are usually called categories (Wang and Strong, 1996), views (Daniel et al., 2008), or hyperdimensions (Karr et al., 2006).

The studies found do not only differ regarding the level of detail at which quality is considered, but also with respect to the part of the statistical process to which they refer. This is discussed in the next paragraph.

1.1.2. Input, throughput (processing), and output quality

The statistical process at NSI’s can roughly be divided in three consecutive phases:

- The input phase: in which input is obtained for the statistical process;
- The throughput (or processing) phase: in which the input is processed to statistical intermediate products;
- The output phase: in which statistical intermediate products are converted to statistical output (tables or micro data).

For these three phases often different quality indicators are used (Vale, 2008; Cerroni et al., 2010). Since we are interested in the quality of secondary data sources to be used for statistics our focus is on the *input phase*. However, not every study distinguishes these phases very clearly.

Given this background we now continue by summarizing the main results of the performed literature review on the composition of quality according to NSI’s (chapter 3) and according to other research areas (chapter 4). Since almost all of the papers found referred to quality at the dimensional level, we will use this level for comparison. However, because of the fact that the definition of these dimensions is not always identical (Batini et al., 2009) we will also translate them to the characteristics listed in Table 1.

2. INPUT QUALITY CHARACTERISTICS CONSIDERED BY NSI’s AND OTHER INSTITUTES

In trying to identify the input quality characteristics considered important by NSI’s and other institutes, we found that they tend to use one of two approaches. In one approach NSI’s study quality by distinguishing several characteristics of data and consecutively distinguishing several dimensions of data that should be looked upon. In the alternative approach NSI’s analyse the data collection and processing process and determine for every step in this process which errors can occur or can be introduced (Ruddock, 1998). As obtaining input is part of the statistical process this approach can also point at “input data quality” issues. We consider both approaches in this chapter.

2.1. Dimensions and characteristics of data quality

In this paragraph we discuss the approach in which data quality dimensions are identified. In doing so, we successively consider primary data and secondary data. We include our findings on primary data, i.e. data collected by the NSI’s themselves, since a lot of research is performed regarding quality of primary data (Daas et al., 2008). This research can give important insights into input data characteristics that should be considered for secondary data.

2.1.1. Primary sources (survey data)

A rich literature exists on the topic of survey data quality, see Groves et al. (2009), Biemer and Lyberg (2003), and Kalton (2001) and the references therein. Definitions of the concept of survey data quality proliferate somewhat, but cluster around the idea that the characteristics of the data collected meet or exceed the stated or implied needs of the user. Several of the above mentioned authors have suggested breaking down the quality of survey data into components or characteristics that focus around the key concepts of:

> 1) Relevance, 2) Accuracy, 3) Timeliness, 4) Accessibility, 5) Interpretability, and 6) Coherence.

When we compare this list with the characteristics in Table 1 and the dimensions of quality included in our checklist regarding the Metadata domain (Daas et al., 2009), we can roughly state that ‘(1) Relevance’ is already included in the metadata checklist. The same holds for the part of ‘(5) Interpretability’ referring to the question: “is the source normally delivered with understandable metadata”. The question “do understandable metadata accompany *this particular* delivery” can in fact be seen as a technical check before testing the quality of the input data in detail.

The other dimensions do belong to the data domain and do roughly correspond to the characteristics: *completeness, correctness, selectivity, timeliness*, and *coherence*. The characteristic *selectivity* can be considered part of the dimension Accuracy. Accessibility of the data can be interpreted as *completeness* but also as a *technical check* in which the availability of all the data in the file is determined (Daas et al., 2008; 2010). We will use the latter interpretation in this paper and include it as an additional characteristic from hereon.

2.1.2. Secondary sources (administrative data)

There are only a relatively small number of studies that specifically focus on the quality aspects of secondary data used for statistical purposes (Daas et al., 2008). The most important papers and books in this area are: Wallgren and Wallgren (2007), Eurostat (2003b), Karr et al. (2006), Unece (2007), Thomas (2005), ONS (2005), and Vale (2008). As one of the few NSI’s, Statistics Netherlands has put considerable effort in developing a quality framework for secondary data sources (Daas et al., 2008). In it the determination of the metadata and data related quality aspects of secondary data sources are clearly separated (Daas et al., 2010). At the highest level, the framework is composed of three hyperdimensions, which are called Source, Metadata, and Data (Daas et al., 2008). The combined set of indicators in the Source and Metadata hyperdimensions contains all quality indicators specific to the Metadata *domain* of quality (Daas et al., 2009). The quality indicators for the Data domain are included in the Data hyperdimension. Since this is the topic of current research any proposals for the content of this hyperdimension are not included here.

2.2. Quality issues derived from analysing errors in the data collection process

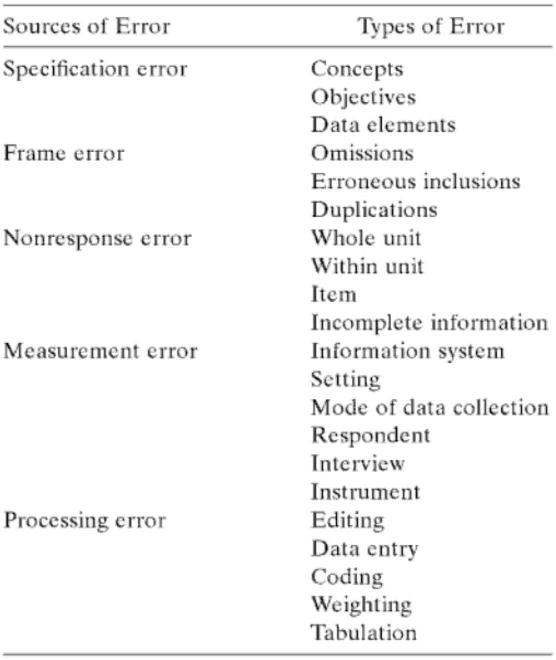
In this paragraph approaches based on an analysis of the sources of error in the data collection process are discussed. This approach tends to focus on quality aspects belonging to the accuracy dimension (Batini and Scannapieco, 2006).

2.2.1. Primary data errors (survey data)

At the highest level, the total error in survey data is composed of sampling and non-sampling errors. A sampling error is the result of the uncertainty associated with an estimate that is based on data gathered from a sample of the population rather than the full population. Different samples will very likely produce (slightly) different estimates. The two major causes of sampling error are errors made in drawing samples and errors made in the estimation process (Bethlehem, 2009). Since the use of secondary data does not include drawing samples such errors are ignored in the remainder of this paper.

Non-sampling errors are the other types of error that affect a survey estimate apart from sampling error (Biemer and Lyberg, 2003). The major types of non-sampling error discerned are: Specification error, Frame error, Nonresponse error, Measurement error, and Processing error (Table 2). The errors being classified as Nonresponse errors and Measurement errors can give insights into the input data quality characteristics that need to be included in our framework. These errors correspond to the following data quality characteristics: *authenticity*, *completeness*, *correctness*, *coherence*, and *selectivity*.

Table 2. Five major sources of non-sampling error and their potential causes (from Biemer and Lyberg, 2003)



Sources of Error	Types of Error
Specification error	Concepts
	Objectives
	Data elements
Frame error	Omissions
	Erroneous inclusions
	Duplications
Nonresponse error	Whole unit
	Within unit
	Item
	Incomplete information
Measurement error	Information system
	Setting
	Mode of data collection
	Respondent
	Interview
	Instrument
Processing error	Editing
	Data entry
	Coding
	Weighting
	Tabulation

Specification errors, Frame errors and Processing errors can be made by the data source keeper resulting in, for example, multiple registrations of the same object or the absence of objects. Such errors do influence input data quality. These errors will influence the *completeness, correctness, coherence,* and *selectivity* characteristics of the data.

2.2.2. Secondary data errors (administrative data)

When the causes of error in the statistical use of secondary data sources are studied (Bakker, 2010) three things stand out. The first one is that, when the secondary data source covers the whole population, the only causes of error are the Non-sampling errors. The second observation is that several new causes of non-sampling are introduced. Examples of this are Linkage and Correction errors (Figure 2). Errors occurring before the NSI obtains the data need to be considered because they affect the quality of the input data. The third observation is that the Nonresponse error must be renamed to Missing data error (Zhang, 2010); an alternative name could be Nonreporting error. Response is a concept that is not very applicable to a secondary data source. All of these changes are the result of the secondary nature of the data in the source. Because the NSI was not involved in data collection, data maintenance, and metadata definition new problem areas arise.

Bakker (2010) gives an overview of the different sources of errors and their location in the process of the use of secondary data (Figure 2). The fact that new sources of error are introduced when secondary data sources are used for statistics, clearly demonstrates that the way NSI's have been looking at the quality of survey data is quite specific for primary data collection. Clearly, more research is needed in this area for secondary data collection.

The errors identified in figure 2 correspond to the following data quality characteristics: *authenticity, completeness, correctness, coherence, selectivity*, and *uniqueness*.

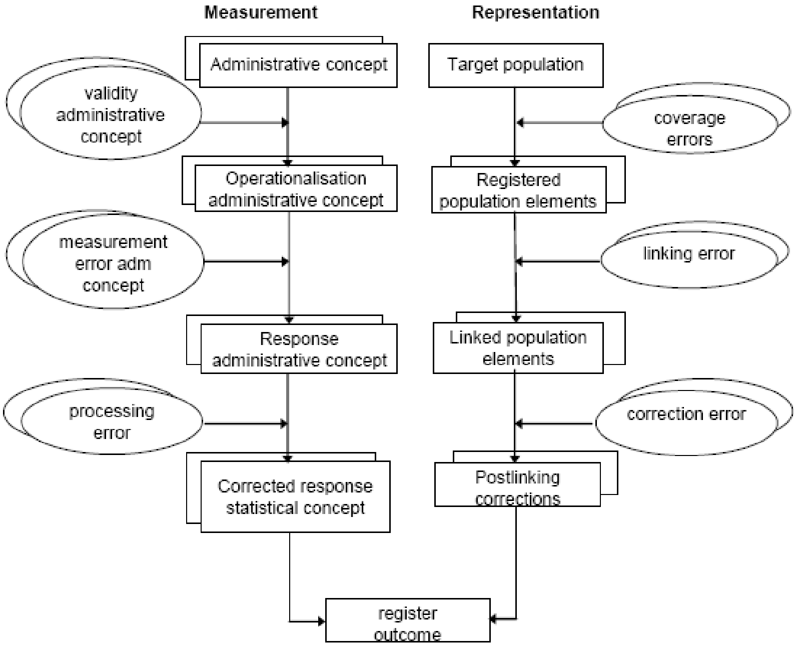

Measurement
Representation
Administrative concept
Target population
validity administrative concept
coverage errors
Operationalisation administrative concept
Registered population elements
measurement error adm concept
linking error
Response administrative concept
Linked population elements
processing error
correction error
Corrected response statistical concept
Postlinking corrections
register outcome

Figure 2. Sources of error in a combined register-approach (from Bakker, 2010).

3. OVERVIEW OF THE STUDY OF DATA QUALITY IN OTHER RESEARCH AREAS

In the previous chapter we considered the approaches used for identifying data quality at NSI's. In this section an overview is given of (data) quality studies performed by other institutes in research areas other than that of statistics. More specific the following research areas are considered:

- Information sciences, section 4.1
- Medical and biosciences, section 4.2
- Social sciences, section 4.3
- Quality research, section 4.4

The authors have tried to be as complete as possible but, because of the broadness of this field of research and the limited time available, there is always a chance that some publications were missed. On the other hand, not all papers found are included in this document. For some research areas, such as econometrics and psychology, data quality studies did not differ considerably from those used by other areas, such as sociology. Since this approach was already discussed, the former approaches were deliberately ignored.

3.1. Information sciences

In the field of the information sciences data quality is usually named 'information' quality. According to Batini et al. (2009) and Knight and Burn (2005), the terms are interchangeable. However, not everybody agrees on this. Some insist on a distinction between data quality and information quality (Churchman, 1971). This distinction would be akin to the distinction between syntax and semantics where for example, the semantic value of 'one' could be expressed in different syntaxes like 00001, 1.0000, 01.0, or 1. Thus a data difference may not necessarily represent poor information quality. Despite of this distinction, it is however clear that data and information quality are related.

3.1.1. General information point of view

In the information sciences information quality is a measure of the usability of the information for the user of that information. Thus in our terms a measure for the quality of the input data. Information quality encompasses many dimensions. A list of categories and dimensions used in assessing the quality of information is proposed by Wang and Strong (1996). The list is shown in Table 3.

Table 3. Categories and dimensions used in assessing quality of information

Categories	*Dimensions*
Intrinsic IQ*	Accuracy, Objectivity, Believability, Reputation
Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Appropriate Amount of Information
Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent representation
Accessibility IQ	Accessibility, Access security

* IQ = Information Quality

When we compare this list with the characteristics in Table 1 we can generally state that 'Ease of understanding', 'Interpretability', 'Relevancy', 'Value-added', and 'Access security' are part of the metadata quality domain. Most of the other dimensions refer to the characteristics: *correctness,* *timeliness, detailedness,* and *completeness.* Note furthermore that again 'accessibility' is listed as important. This is of course not surprising as inaccessible data are useless.

3.1.2. Information from web pages

In the information sciences the quality of data available on web pages is also studied a lot. The paper of Knight and Burn (2005) provides an excellent overview of the state of art and advancements made in this area. Both data and metadata quality aspects are studied and attempts are made to include the users' point of view on the quality of the information collected. Many different frameworks have been developed to determine the quality of information on the internet; particularly the World Wide Web. Knight and Burn (2005) provide and overview of the twenty most commonly used dimensions for quality in this area. These are, in decreasing order of importance:

> 1) Accuracy, 2) Consistency, 3) Security, 4) Timeliness, 5) Completeness, 6) Concise(ness), 7) Reliability, 8) Accessibility, 9) Availability, 10) Objectivity, 11) Relevancy, 12) Useability, 13) Understandability, 14) Amount of data, 15) Believability, 16) Navigation, 17) Reputation, 18) Useful(ness), 19) Efficiency, and 20) Value-added.

Several of these dimensions refer to metadata quality, these are: (3) Security, (11) Relevancy, (12) Useability, (13) Understandability, (18) Useful(ness), (19) Efficiency, and (20) Value-added. The dimension 'Navigation' is typical for the information sciences and not of interest for us. The other dimensions direct or indirectly refer to the following characteristics: *correctness, coherence, timeliness,* and *completeness.* Here, again the technical check 'accessibility' is included.

More details can be found in the paper of Eppler and Muenzenmayer (2002) that describes the measurement of a lot of these dimensions in more detail.

3.2. Medical and biosciences

In this section we distinguish between epidemiology (paragraph 4.2.1), the study of medical registrations (paragraph 4.2.2), and biology (paragraph 4.2.3).

3.2.1. Epidemiology

In epidemiological research quite some secondary data sources are used as input. To cope with the quality issues of the data in these sources, Sørensen et al. (1996) developed a quality framework for its evaluation. The 'factors' included in the framework are:

> 1) Completeness of registration of individuals, 2) Accuracy and degree of completeness of the data, 3) Size of the data source, 4) Registration period, 5) Data accessibility, availability, and costs, 6) Data format, and 7) Record linkage.

Methods of determination of each of those 'factors' are discussed in the corresponding paper. Note that the list contains 'factors' belonging to the data and metadata domain of quality. The factors referring to the data domain are 1, 2, 4, and 7. The first three correspond to the characteristics: *completeness, correctness,* and *timeliness.* Number 7, the extent to which data can be linked, refers to the characteristic *uniqueness.* Here also some technical checks, viz. "data accessibilty" and "data format", are listed.

3.2.2. Medical registrations

A lot of effort is put into the evaluation of the quality of the data in medical registrations. Cancer registries are the sources most commonly studied (Skeet, 1991). The papers of Bray and Parkin provide a recent overview of the dimensions and methods used for the determination of the quality of the data in medical registries (Bray and Parkin, 2009; Parkin and Bray, 2009). Usually the information in a random selection of patient records (on paper) is compared with the information stored in the register for those patients. The information in the patient records is considered 'the golden standard'. The dimensions of data quality discerned in this field of research are:

1) Comparability, 2) Validity (accuracy), 3) Timeliness, and 4) Completeness.

These dimensions correspond to the Table 1- characteristics: *coherence*, *correctness*, *timeliness*, and *completeness*. Larsen et al. (2009) applied these findings to the Norwegian cancer register.

4.2.3 Biology

Stribling et al. (2003) define four 'performance characteristics' to document the data quality of taxonomies. Some of these characteristics resemble dimensions. They are:

1) Accuracy, 2) Precision, 3) Bias, and 4) Completeness.

This implies that according to this study the following Table 1- characteristics seem to be important: *correctness, detailedness, selectivity,* and *completeness.*

3.3. Social sciences

In the social sciences we distinguish between sociology (section 4.3.1) and history (section 4.3.2).

3.3.1. Sociology

In sociology surveys are the main data source used as input. The quality of these input data is often assessed by seven dimensions (Biemer and Lyberg, 2003). These are:

1) Relevance, 2) Accuracy (composed of the total survey error), 3) Timeliness, 4) Accessibility, 5) Comparability, 6) Coherence, and 7) Completeness.

The dimension 'Relevance' in this list refers to metadata quality aspects. The remaining dimensions do roughly stated correspond to the following Table 1- characteristics: *correctness, selectivity, timeliness, coherence,* and *completeness*. Here, again the technical check 'accessibility' is included.

3.3.2. History

Because of the secondary nature of most of the sources used by historians, researchers in this field have always devoted a lot of time on a thorough review of the quality of the metadata of the sources used. Contextual information and objectivity of the information in the source is considered a very important topic (Howell and Prevenier (2001). This suggests the Table 1- characteristic: *correctness.*

To illustrate the great diversity of data sources used by historians a few examples are given, such as: Roman text on papyrus, Egyptian statues, medieval fabric, and pictures and movies of the first and second world war. Despite the fact that secondary sources are the predominant sources of information for almost all historians, with the exception of scholars that study very recent history, no general framework could be found for determining the quality research of those sources. This is probably caused by the great variety in the sources used. A thorough overview of the use of secondary sources by historians, and some of the methods used to verify them, is provided by Howell and Prevenier (2001).

3.4. Quality research

The research areas described in the previous sections of this chapter all shared the common characteristic that determining the quality of input data was required for performing the actual research. There is however also a research area in which defining and determining quality of data is the main topic of research. This area of research is discussed in this section. The four most important studies are discussed.

3.4.1. Results of Wang et al.

Wang et al. (1995) have studied the dimensional composition of quality. For clarity, this concerned both the data and metadata domains of quality. The most important finding of Wang et al. (1995) was that no general agreement exists on the number and 'types' of dimensions discerned. However, regarding the dimensions they concluded that the quality dimensions most frequently mentioned were:

> 1) Accuracy, 2) Timeliness, 3) Completeness, and 4) Consistency.

In terms of Table 1- characteristics this corresponds to: *correctness*, *coherence, completeness, timeliness*, and *selectivity*.

3.4.2. Results of Wand and Wang

A more extensive list of the most noted quality dimensions studied by some of the previous authors is published in Wand and Wang (1996) and shown in table 4.

Table 4. Most notable dimensions of quality (from Wand and Wang, 1996)

Dimensions	No. cited	Dimensions	No. cited	Dimensions	No. cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

In the first ten dimensions, the characteristics *correctness, selectivity, timeliness, completeness*, and *coherence* are found. Notice that both data and metadata dimensions of quality are listed in table 4.

3.4.3. Results of Batini et al.

Batini et al. (2009) followed a similar approach as Wang and co-workers but specifically focussed on the quality of the data. This study confirmed the observation of Wang et al. (1995) that no agreement consists on the set of dimensions to be used for data quality. According to Batini et al. (2009) the most commonly used dimensions for data quality are:

> 1) Accuracy, 2) Completeness, 3) Consistency, and 4) Time-related dimensions (Currency, Volatility, and Timeliness).

These dimensions correspond to the characteristics: *correctness, completeness, coherence, stability,* and *timeliness*. Another important finding of Batini et al. (2009) was that he noticed that the exact definitions of the dimensions used (even the common ones) varied between studies.

3.4.4. Results of Redman

The data quality expert Thomas C. Redman (2001) has also created his own set of dimensions for data quality (Table 5). These were derived from practice and are divided into two categories. In first category of this set the Table 1- characteristics: *correctness, selectivity, timeliness, completeness,* and *coherence* are mentioned. The technical check accessibility is also observed here. In the second category the characteristic *detailedness* is included.

Table 5. Categories and dimensions of data quality according to Redman (2001).

Categories	*Dimensions*
Free of defects	Accessible, Accurate, Timely, Complete, Consistent with other sources
Possesses desired features	Relevant, Comprehensive, Proper level of detail, Easy to read, Easy to interpret

4. CONCLUSIONS

The aim of this paper was to identify the most important characteristics of input data quality. These should be included in our framework. We therefore considered which characteristics of input data quality are distinguished in several research areas that use secondary data sources. The characteristics used are shown in Table 1 (page 5) and were specifically identified for data quality in a previous study by one of the authors (Daas and Van Nederpelt (2010). By counting the number of times a characteristic has been identified in our literature study, the most important ones can be objectively identified. The results of this are shown in Table 6.

The characteristics *coherence, completeness, correctness, selectivity*, and *timeliness* were mentioned the most (Table 6). It is clear that these characteristics are very important when studying data quality in various sources. In addition, this study also revealed that the '*accessibility*' of the data, from a technical point of view, is important. This suggests that some 'technical checks' also need to be included as part of the essential input data quality characteristics.

From table 6 it is clear that the characteristics: *authenticity, confidentiality, detailedness, stability,* and *uniqueness* were hardly mentioned. This suggests to exclude these characteristics in the set of properties that need to be studied for a data source. This does of course not mean that the characteristics mentioned only seldom are not important in determining input data quality. For example, when the aim of an NSI is to produce time series for an important statistic and the required input data can only be obtained from an external data source *stability* is very important. However when a researcher uses a source only once *stability* is not a key issue.

Consequence for workpackage 4 (WP4) of the BLUE-ETS project is that quality indicators covering the essential characteristics mentioned above, must certainly be included in the list of quality indicators that is going to be produced for determining the input quality of the data included in administrative data sources. The big advantage of this finding is that it limits the focus of the work of WP4 to these essential properties of data quality. During the construction of the list of indicators and during tests, this limited focus could be confirmed or, if not, should be (slightly) adjusted. The inclusion of any of the other characteristics could, for instance, highly depend on the intended use of the source by the user.

When the essential characteristics are interpreted from the viewpoint of dimensions of data quality, the following list of proposed dimensions emerges, namely: Coherence, Completeness, Accuracy (the combination of *correctness* and *selectivity*), and Timeliness. As mentioned by Batini et al. (2009) the latter dimension could probably be better named Time-related dimensions to cover all time-related data quality issues. In addition, inclusion of a technical checks dimension should seriously be considered.

Table 6. Total times a characteristic of data quality was mentioned in the studies included in this paper

	Authenticity	Coherence	Completeness	Confidentiality	Correctness	Detailedness	Selectivity	Stability	Timeliness	Uniqueness	Technical check (accesibility)
Survey data (3.1.1)		x	x		x		x		x		x
Secondary data (3.1.2)	?	?	?	?	?	?	?	?	?	?	?
Survey data, errors (3.2.1)	x	x	x		x		x				
Secondary data, errors (3.2.2)	x	x	x		x		x			x	
Information sciences (4.1.1)			x		x	x			x		x
Information sciences, web (4.1.2)		x	x		x				x		x
Epidemiology (4.2.1)			x		x				x	x	x
Medical registrations (4.2.2)		x	x		x				x		
Biology (4.2.3)			x		x	x	x				
Sociology (4.3.1)		x	x		x		x		x		x
History (4.3.2)					x						
Quality research, Wang et al. (4.4.1)		x	x		x		x		x		
Quality research, Wand & Wang (4.4.2)		x	x		x		x		x		
Quality research, Batini et al. (4.4.3)		x	x		x			x	x		
Quality research, Redman (4.4.4)		x	x		x	x	x		x		x
Total times menitoned	2	10	13	0	14	3	8	1	10	2	6

REFERENCES

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.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A. (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys 41(3), Article 16, July.

Batini, C., Scannapieco, M. (2006) Data Quality: Concepts, Methodologies and Techniques. Springer, Berlin, Germany.

Bethlehem. J.G. (2009), Applied Survey Methods, A Statistical Perspective. John Wiley and Sons, Hoboken, USA.

Biemer P.P., Lyberg L.E. (2003) Introduction to Survey Quality, John Wiley and Sons, Hoboken, U.S.A.

Bray, F., Parkin, D.M. (2009) Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. European Journal of Cancer 45, pp. 747-755.

Cerroni, F., Migliardo, S., Morganti E. (2010) Quality evaluation analysis of the Italian business register on enterprise groups. In: Proceedings of Q2010 European Conference on Quality in Official Statistics, Statistics Finland and Eurostat, Helsinki, Finland.

Churchman, C.W. (1971) The design of inquiring systems. Basic Books, New York, U.S.A.

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

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010) Determination of Administrative Data Quality: Recent results and new developments. In: Proceedings of 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.

Daas, P.J.H., Van Nederpelt, P.W.M. (2010) Application of the Object Oriented Quality Management model to Secondary Data Sources. Discussion paper 10012, Statistics Netherlands, the Hague/Heerlen, The Netherlands.

Daniel, F., Casati, F., Palpanas, T., Chayka, O., Cappiello, C. (2008) Enabling Better Decisions through Quality-Aware Reports in Business Intelligence Applications, 13th International Conference on Information Quality 2008, November, Boston, pp. 310-324.

Ehling, M., Körner, T. (2007), Handbook on Data Quality Assessment Methods and Tools. European Commission, Wiesbaden, Germany.

Eppler, M., Muenzenmayer, P. (2002) Measuring information quality in a web context: A survey of state-of-art instruments and an application methodology. In: Proceedings of the 7th International Conference on Information Quality, pp. 187-196.

ESC (2007) Pros and cons for using administrative records in statistical bureaus, paper presented at the seminar on increasing the efficiency and productivity of statistical offices. Economic and Social Council conference of European statisticians, Geneva, Switzerland.

Eurostat (2003a) Item 4.2 Methodological documents Definition of quality in statistics, Working group Assessment of quality in statistics, Eurostat, Luxembourg.

Eurostat (2003b) Item 6 Quality assessments of administrative data for statistical purposes. Working group Assessment of quality in statistics, Eurostat, Luxembourg.

Groves, R.M, Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (2009) Survey Methodology, 2nd edition. Wiley Series in Survey Methodology, Hoboken, U.S.A.

Howell, M., Prevenier, W. (2001) From Reliable Sources, an Introduction to Historical Methods. Cornell University Press, Ithaca, U.S.A.

ISO (1986) ISO 8402, Quality -- vocabulary. International Organization for Standardisation. (Note: This version is replaced by ISO 9000:2005, Quality management systems -- Fundamentals and vocabulary).

Juran, J.M. (2004) Architect of Quality: The Autobiography of Dr. Joseph M. Juran. McGraw-Hill, New York, USA.

Kalton, G. (2001) How Important is Accuracy? In: Proceedings of Symposium on Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada, Quebec, Canada.

Karr, A.F., Sanil, A.P., Banks, D.L. (2006) Data quality: A statistical perspective, Statistical Methodology, 3, pp. 137-173.

Knight, S-A, Burn, J. (2005) Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal 8, pp. 159-172.

Larsen, I.K., Småstuen, M., Johannesen, T.B., Langmark, F., Parkin, D.M., Bray, F., Møller, B. (2009) Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. European Journal of Cancer 45, pp. 1218-1231.

ONS (2005) Guidelines for measuring statistical quality, version 3.0. Office of National Statistics, London, UK.

Parkin, D.M., Bray, F. (2009) Evaluation of data quality in the cancer registry: Principles and methods. Part II: Completeness. European Journal of Cancer 45, pp. 756-764.

Redman, T.C. (2001) Data Quality: The Field Guide. Digital Press, Woburn, U.S.A.

Ruddock, V. (1998) Measuring and Improving Data Quality. Government Statistical Service Methodology (GSSM) Series publication no. 14, Office of National Statistics, London, UK.

Skeet, R.G. (1991) Quality and quality control. In: Cancer registrations, principles and methods, Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S., Skeet, R.G., Eds. IARC Scientific Publications No. 95, Lyon, France, pp. 101-107.

Sørensen, H.T., Sabroe, S., Olsen, S. (1996) A Framework for Evaluation of Secondary Data Sources for Epidemiological Research. International Journal of Epidemiology 25(2), pp. 435-442.

Statistics Finland (2004) Use of Register and Administrative Data Sources for Statistical Purposes. Handbook 45, Statistics Finland, Helsinki, Finland.

Stribling, J.B., Moulton II, S.R., Lester, G.T. (2003) Determining the Quality of Taxonomic Data. Journal of the North American Benthological Society 22(4), pp. 621-631.

Thomas, M. (2005) Assessing Quality of Administrative Data, Survey Methodological Bulletin, 56, pp. 74-84.

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..

Vale, S. (2008) Using Administrative Source for Official Statistics - A Handbook of Principles and Practices, version 1.1, April, United Nations Economic Commission for Europe.

Van Nederpelt, P.W.M. (2009). The creation and application of a new quality management model. Discussion paper 09040, Statistics Netherlands, The Hague/Heerlen.

Wade, J. (2005) Soapbox: on quality. QualityWorld 14, p. 14, August.

Wallgren, A., Wallgren, B. (2007) Register-based Statistics: Administrative Data for Statistical Purposes, Wiley, Chichester, U.S.A.

Wand, Y., Wang, R.Y. (1996) Anchoring Data Quality Dimensions in Ontolological Foundations. Communications of the ACM, 39 (11), pp. 86-95.

Wang, R.Y., Storey, V.C., Firth, C.P. (1995) A framework for analysis of data quality research. IEEE Trans. on Knowl. Data Eng. 7 (4), pp. 623–640.

Wang, R.Y., Strong, D.M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12 (4), pp. 5-34.

Zhang, L-C. (2010) Assessment of uncertainty in register-based small area means of a binary variable. Presentation at the Workshop on Measurement Errors in Administrative Data, Mannheim, Germany, 14-15 June.