Checklist for the Quality evaluation of Administrative Data Sources

Piet Daas, Saskia Ossen, Rachel Vis-Visschers and Judit Arends-Tóth

Discussion paper (09042)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2007–2008	= 2007 to 2008 inclusive
2007/2008	= average of 2007 up to and including 2008
2007/'08	= crop year, financial year, school year etc. beginning in 2007 and ending in 2008
2005/'06–2007/'08	= crop year, financial year, etc. 2005/'06 to 2007/'08 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309042 X-10

CHECKLIST FOR THE QUALITY EVALUATION OF ADMINISTRATIVE DATA SOURCES

Piet Daas, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth

Summary: Statistics Netherlands is increasingly making use of administrative and other secondary data sources for the production of statistics. This approach makes Statistics Netherlands highly dependent on the quality of those sources. It is therefore of vital importance that a procedure is available to determine the quality of such data sources in a systematic, objective, and standardized way. For this purpose a quality framework and a checklist have been developed. The framework distinguishes three different views on quality, namely the Source view, the Metadata view, and the Data view. The Source view focuses on quality aspects essential for the delivery of the data source, whereas the Metadata view focuses on the metadata aspects of the data source. In the Data view technical and accuracy related aspects of data quality are studied. By means of a checklist the quality indicators of the Source and Metadata part of the framework are determined. The Data view is not included. The checklist aims to minimize the effort and time required for evaluation. In this paper the quality framework, the checklist, its application, and the evaluation results obtained are discussed.

Keywords: Administrative data sources, Quality determination, Register-based statistics

Contents

1. Introduction 5
2. Quality of administrative data sources 6
2.1 Quality framework 6
2.1.1 Source 7
2.1.2 Metadata 7
2.1.3 Data 8
2.2 Application of the framework 10
2.2.1 Evaluation sequence 10
2.2.2 Checklist 11
2.3 Source material 11
3. Results and discussion 13
3.1 Application 13
3.2 Scores obtained 13
3.3 Source discussion 14
3.4 Metadata discussion 14
3.5 Use of the checklist 15
4. Concluding remarks 16
References 17
Annex, the checklist 19

1. Introduction

National Statistical Institutes (NSI's) need data for the production of statistics. Apart from data obtained through surveys, NSI's are increasingly using data collected and maintained by other, non-statistical, organizations. Administrative data is an example of such a data source (Wallgren and Wallgren, 2007). It is produced as a result of administrative processes of organizations but it is -very often- also an interesting data source for NSI's. During the last decade, more and more NSI's have realized this (Unece, 2007). A major advantage of using administrative data for statistics compared to survey data is that it reduces the costs of data collection and reduces the administrative burden on enterprises and persons. Since administrative data often covers whole populations, it is also very well suited for creating detailed and longitudinal statistics on subpopulations and regions (Wallgren and Wallgren, 2007).

From a statistical point of view, administrative data also has some disadvantages. For example, the collection and processing of administrative data is beyond the control of the NSI. It is the data source keeper who manages these aspects, and not the NSI. The same is true for the units and variables an administrative data source contains. These are defined by administrative rules and may therefore not be identical to those required by an NSI (Wallgren and Wallgren, 2007). The disadvantages are predominantly the result of the fact that, in most cases, an NSI uses an administrative data source for a purpose different than the one for which the data was originally collected. As a result of this difference, the 'statistical' usability of a data source needs to be thoroughly studied by an NSI prior to its use. This often takes considerable effort (Bakker, 2009; ESC, 2007; Van der Laan, 2000). Since NSI's want to produce high quality statistics (Statistics Netherlands, 2008), which are affected by the quality of the input data, it is of vital importance that NSI's are able to determine the quality of administrative data sources in an efficient and standardized way. For this purpose a quality framework was developed by Statistics Netherlands.

The quality framework enables the determination of the quality of secondary data sources, such as administrative data sources (Daas et al., 2008b). It is also embedded in the new quality management model of Statistics Netherlands (Van Nederpelt, 2009). The framework contains the total of the quality aspects identified for administrative data sources by Statistics Netherlands (Daas and Fonville, 2007) and those mentioned in publications by others. Readers are referred to the papers of Daas et al. (2008a-b) for a more detailed description of the identification and combination of these aspects. The present paper focuses on the way the framework should be applied.

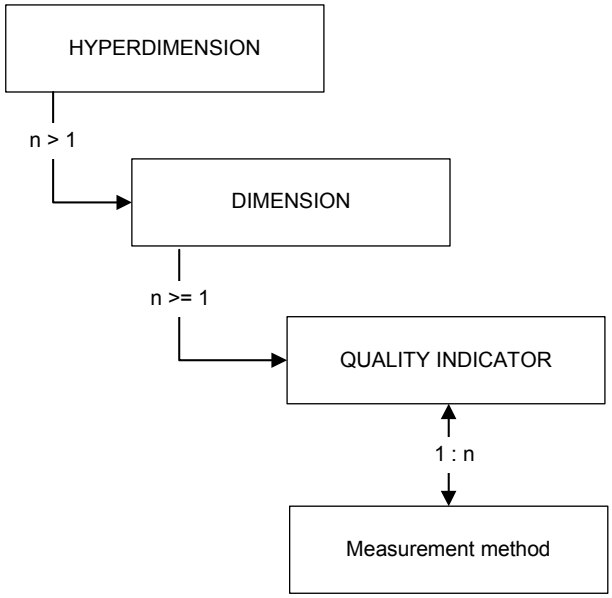
2. Quality of administrative data sources

2.1 Quality framework

The quality framework for administrative data sources is composed of several high level views on the quality of a data source. In the literature these are also called categories (Batini and Scannapieco, 2006) or hyperdimensions (Karr et al., 2006). The latter term will be used in the remainder of this paper. The quality aspects in each hyperdimension influence the usability of a data source in a different way. Three hyperdimensions, i.e. Source, Metadata, and Data, are used to determine the statistical usability of an administrative data source (Daas et al., 2008b). Each hyperdimension is composed of several dimensions; each dimension contains a number of quality indicators (Figure 1). A quality indicator is measured or estimated by one or more either qualitative or quantitative methods (Daas et al., 2008a-b).

The Source, Metadata, and Data hyperdimension each highlight different quality aspects of a data source. The hyperdimensions are also ordered according to an increasing level of detail. The quality indicators in the Data hyperdimension, for instance, report on quality aspects at a much more detailed level than the quality indicators included in the Metadata hyperdimension. The same is true for the Metadata and Source hyperdimensions. An important result of this ordered distinction is the fact that it efficiently guides the user in the study of the quality of a data source. The sequential study of each hyperdimension prevents that the user invests time and effort in the determination of quality aspects not (yet) relevant at that point in time. Next we give an overview of the dimensions, quality indicators, and measurement methods for the Source, Metadata, and Data hyperdimensions.

Figure 1. Hierarchical relation between the different aspects of quality used in the framework developed



2.1.1 Source

In the Source hyperdimension, the quality aspects related to the data source as a whole, the data source keeper, and the delivery of the data source to the NSI are studied. The Source hyperdimension is composed of five quality dimensions; these are: Supplier, Relevance, Privacy and security, Delivery, and Procedures. In table 1 the dimensions, quality indicators, and measurement methods for the Source hyperdimension are listed. In the Source hyperdimension mainly qualitative methods are present. Exceptions are the calculations of the effect of the use of the data source on i) the administrative burden induced by the NSI and on ii) the costs of the NSI.

2.1.2 Metadata

The Metadata hyperdimension specifically focuses on the metadata related aspects of the data source. Clarity of the definitions and completeness of the meta information are some of the quality aspects included. The Metadata hyperdimension is composed of four dimensions: Clarity, Comparability, Unique keys, and Data treatment (by the data source keeper). The Data treatment dimension is a special case. It consists of quality indicators used to determine whether the data source keeper performs any checks on and/or modifies the data in the source. This meta-

Table 1. Dimensions, quality indicators, and methods for Source

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Supplier	1.1 Contact	-Name of the data source -Data source contact information -NSI contact person
	1.2 Purpose	-Reason for use of the data source by NSI
2. Relevance	2.1 Usefulness	-Importance of data source for NSI
	2.2 Envisaged use	-Potential statistical use of data source
	2.3 Information demand	-Does the data source satisfy information demand?
	2.4 Response burden	-Effect of data source use on response burden
3. Privacy and security		
	3.1 Legal provision	-Basis for existence of data source
	3.2 Confidentiality	-Does the Personal Data Protection Act apply? -Has use of data source been reported by NSI?
	3.3 Security	-Manner in which the data source is send to NSI -Are security measures required? (hard/software)
4. Delivery	4.1 Costs	-Costs of using the data source
	4.2 Arrangements	-Are the terms of delivery documented? -Frequency of deliveries
	4.3 Punctuality	-How punctual can the data source be delivered? -Rate at which exceptions are reported -Rate at which data is stored by data source keeper
	4.4 Format	-Formats in which the data can be delivered
	4.5 Selection	-What data can be delivered? -Does this comply with the requirements of NSI?
5. Procedures	5.1 Data collection	-Familiarity with the way the data is collected
	5.2 Planned changes	-Familiarity with planned changes of data source -Ways to communicate changes to NSI
	5.3 Feedback	-Contact data source keeper in case of trouble? -In which cases and why?
	5.4 Fall-back scenario	-Dependency risk of NSI -Emergency measures when data source is not delivered according to arrangements made

information is very important for an NSI as it certainly affects the quality of the product delivered by the data source keeper. In table 2 all the dimensions, quality indicators, and measurement methods are shown for the Metadata hyperdimension. The Metadata hyperdimension solely contains qualitative methods.

2.1.3 Data

The Data hyperdimension focuses on the quality aspects of the data (facts) in the data source. Although the majority of the results described in this paper focus on the quality aspects included in the Source and Metadata hyperdimension, the Data hyperdimension is discussed here for completeness sake.

The quality aspects of the Data hyperdimension are predominantly accuracy related with the exception of those included in the Technical Checks dimension (table 3). This dimension contains indicators that verify the readability of the data file and the compliance of the data to the metadata definition. The other nine, accuracy related, quality dimensions of the Data hyperdimension are: Over coverage, Under coverage, Linkability, Unit non response, Item non response, Measurement, Processing, Precision, and Sensitivity. The Sensitivity dimension is mainly used to determine the effect of time-dependent changes in the population composition on data quality (Daas et al., 2008b). A considerable part of the measurement methods in the Data hyperdimension are based on the so-called Representativity index (R-index; see

Table 2. Dimensions, quality indicators, and methods for Metadata

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Clarity	1.1 Population unit definition	-Clarity score of the definition
	1.2 Classification variable definition	-Clarity score of the definition
	1.3 Count variable definition	-Clarity score of the definition
	1.4 Time dimensions	-Clarity score of the definition
	1.5 Definition changes	-Familiarity with occurred changes
2. Comparability	2.1 Population unit definition comparison	-Comparability with NSI definition
	2.2 Classification variable definition comparison	-Comparability with NSI definition
	2.3 Count variable definition comparison	-Comparability with NSI definition
	2.4 Time differences	-Comparability with NSI reporting periods
3. Unique keys	3.1 Identification keys	-Presence of unique keys -Comparability with unique keys used by NSI
	3.2 Unique combinations of variables	-Presence of useful combinations of variables
4. Data treatment (by data source keeper)		
	4.1 Checks	-Population unit checks performed -Variable checks performed -Combinations of variables checked -Extreme value checks
	4.2 Modifications	-Familiarity with data modifications -Are modified values marked and how? -Familiarity with default values used

table 3). An R-index, a concept developed by Statistics Netherlands (Schouten and Cobben, 2007), measures the extent to which the composition of the units in a data source, at a certain point in time, deviate from the population. This is important for administrative data sources because the composition of the units present in the data source may be time-dependent; an aspect that is particularly important for administrative data sources used in short-term statistics (Van Delden and Aelen, 2008). Because of the fact that the other time-related data quality issues are covered by R-indices, timeliness is not included as a separate dimension in the Data hyperdimension. Because the quality aspects in the Data hyperdimension are the topic of current research in our office, it is possible that the composition and structure of this hyperdimension might change to the one shown in table 3 of this paper.

Table 3. Dimensions, quality indicators, and methods for Data

DIMENSIONS	QUALITY INDICATORS	METHODS
1. Technical checks	1.1 Readability	-Can all the data in the source be accessed?
	1.2 Metadata compliance	-Does the data comply to the metadata definition? -If not, report the anomalies
2. Over coverage	2.1 Non-population units	-Percentage of units not belonging to population
3. Under coverage	3.1 Missing units	-Percentage of units missing from the target population
	3.2 Selectivity	-R-index [1)] for unit composition
	3.3 Effect on average	-Maximum bias of average for core variable -Maximum RMSE [2)] of average for core variable
4. Linkability	4.1 Linkable units	-Percentage of units linked unambiguously
	4.2 Mismatches	-Percentage of units incorrectly linked
	4.3 Selectivity	-R-index for composition of units linked
	4.4 Effect on average	-Maximum bias of average for core variable -Maximum RMSE of average for core variable
5. Unit non response	5.1 Units without data	-Percentage of units with all data missing
	5.2 Selectivity	-R-index for unit composition
	5.3 Effect on average	-Maximum bias of average for core variable -Maximum RMSE of average for core variable
6. Item non response	6.1 Missing values	-Percentage of cells with missing values
	6.2 Selectivity	-R-index for variable composition
	6.3 Effect on average	-Maximum bias of average for variable -Maximum RMSE of average for variable
7. Measurement	7.1 External check	-Has an audit or parallel test been performed? -Has the input procedure been tested?
	7.2 Incompatible records	-Fraction of fields with violated edit rules
	7.3 Measurement error	-Size of the bias (relative measurement error)
8. Processing	8.1 Adjustments	-Fraction of fields adjusted (edited)
	8.2 Imputation	-Fraction of fields imputed
	8.3 Outliers	-Fraction of fields corrected for outliers
9. Precision	9.1 Standard error	-Mean square error for core variable
10. Sensitivity	10.1 Missing values	-Total percentage of empty cells
	10.2 Selectivity	-R-index for composition of totals
	10.3 Effect on totals	-Maximum bias of totals -Maximum RMSE of totals

[1] R-index: Representative Index, an indicator that estimates the selectivity of the data missing by using information available in other sources (Schouten and Cobben 2007, Cobben and Schouten 2008).
[2] RMSE: root mean square error; a common used statistical measure for the quality of an estimator. The RMSE is equal to the square root of the sum of the bias and variance of the estimator.

2.2 Application of the framework

2.2.1 Evaluation sequence

The framework introduced above is used for the determination of the quality of administrative and other secondary data sources. The quality is determined by successively evaluating the quality aspects included in the Source, Metadata, and Data hyperdimension. This strict order is the result of the fact that the quality aspects in the Source hyperdimension report on quality at a much more general level than the aspects included in the Metadata and Data hyperdimensions. The same is true for the quality aspects of the Metadata hyperdimension in comparison to those of the Data hyperdimension. As a consequence, the user must first evaluate the quality indicators in the Source hyperdimension, then those in the Metadata hyperdimension, and finally those in the Data hyperdimension. This approach prevents that the user invests considerable time and effort in the study of quality aspects that are not relevant at that specific point in time.

When the results for some of the quality indicators in a hyperdimension reveal problems, it is recommended to sort these out before the start of the evaluation of the next hyperdimension. This approach is advised because it prevents that problems observed earlier on in the evaluation are (later on) found to be so sever that they block the use of the data source for the statistical application the user had in mind. When unsolvable problems occur during the evaluation of the Source hyperdimension it is likely that the user has to conclude that the data source cannot be used for statistics at all. When the user has another (new) statistical use in mind for a data source that has already been evaluated, the same sequence of events must be repeated. It is very likely, however, that the results obtained for the Source hyperdimension will not differ a lot from those previously obtained. The quality aspects of the Metadata and Data hyperdimension should always be (re)evaluated for such data sources.

If the evaluation of the last hyperdimension, Data, is successful, the data source can be used for the production of statistics. It is conceivable, however, that the user would like to perform one or more additional -very specific- checks after the evaluation of the three hyperdimensions (Kuijvenhoven and Schouten, 2008). These additional checks will occur at the data level. An example of a specific check is the comparison of the estimated percentage of unemployed persons obtained, after editing and weighting, from an administrative data source (such as the job-seeker information in the register of the Centre for Work and Income) with that of the estimated percentage obtained through the Labour Force Survey of Statistics Netherlands. These types of checks are not included in the quality framework because the framework only contains general applicable quality indicators (Daas et al. 2008b).

2.2.2 Checklist

For the evaluation of the Source and Metadata hyperdimension, the authors have developed a checklist (Daas et al., 2008b; Daas et al., 2009). The checklist guides the user through the quality indicators that need to be evaluated for both Source and Metadata. For the Data hyperdimension a checklist cannot be used because of the large amount of calculations that need to be performed. The best approach for this hyperdimension is the topic of current research. Because of this, the quality aspects in the Data hyperdimension were not determined for the data sources described in this paper. The Source and Metadata checklist can be used for a data source that is already available (and used) by the NSI and for the evaluation of a new data source that could potentially be used for statistics. The checklist is included in the Annex.

The checklist guides the user through the measurement methods for each of the quality indicators shown in tables 1 and 2. By answering the questions in the checklist, the 'value' of every measurement method in tables 1 and 2 is determined. Since the predominant part of the methods in the Source and Metadata hyperdimension are qualitative, usually a score has to be filled in. When problems are found or a question cannot be answered completely, the user is guided in the steps to take. Apart from this, additional space is included to write down remarks. Evaluation of the Metadata-part requires that the user has a particular use in mind (Daas et al., 2008b).

2.3 Source material

To test the usability of the checklist and its usefulness for statistics, six administrative data sources were evaluated. The administrative data sources studied were: Policy record Administration (PA), Student Finance Register (SFR), register of the Centre for Work and Income (CWI), Exam Results Register (ERR), the coordinated register for Higher Education (1FigHE), and the coordinated register for Secondary General Education (1FigSGE). Each of these data sources is described in more detail below.

The PA is maintained by the Institute for Employee Benefit Schemes; a self governing body that works under authority of the Ministry of Social Affairs and Employment. In the PA, all Dutch employers, (ex)employees, and their labour relations are registered. The employee population is that of all insured employees in the Netherlands. The PA is considered one of the largest administrations in the Netherlands; millions of entries are processed every month. The total number of records is about 20 million. Data collection started in 2006 and suffered some start-up problems in the beginning. The PA is a very important register for Statistics Netherlands because it provides, among others things, very detailed information on jobs and the number of jobs in the Netherlands.

The SFR is the registration of study grants in the Netherlands. It is maintained by the Information Management Group. From 1995 onwards students are included. The register contains information on all students receiving a study grant in higher education and on students of 18 years and older with a grant in secondary vocational

education. The number of students registered at least once is 2.1 million (Bakker et al., 2008). The SFR is, among other things, used in educational and income statistics.

The CWI-register contains information on job-seekers in the Netherlands. As of the beginning of this year it is maintained by the Institute for Employee Benefit Schemes. The register contains information on the (previous) jobs, education, and courses of job-seekers. Information is supplied to Statistics Netherlands from 1990 onwards. For more than 5 million people at least one level of education is registered in this source (Bakker et al., 2008). The CWI provides information that is used for the labour statistics and is being studied for use in educational statistics.

The ERR is a register in which all pupils sitting final exams in secondary general education from 1998/'99 onwards are included. It is maintained by the Information Management Group. In the ERR the level of education and the exam results of approximately 1.3 million persons are found. Its use for educational statistics is a topic of discussion (Bakker et al., 2008). Due to a recent change in legislation the ERR now only includes information about students in a very limited number of schools. Information on other students is transferred to the coordinated register on secondary education. See the 1FigSGE below.

The 1FigHE is a register with information on higher education in the Netherlands. The register is based on the Central Register of Higher Education Enrolment maintained by the Information Management Group. The 1FigHE is a harmonized register created by the joint effort of the Ministry of Education, Culture, and Science, the Higher Professional Education Council, the Association of Universities in the Netherlands, and Statistics Netherlands. Standardized variables and derivation rules are used meaning that all cooperating institutions use the same variable definitions and derivation rules. Information from the study year 1985/'86 onwards is available (Bakker et al., 2008; Ossen and Daas, 2009). The source is used for educational statistics.

The 1FigSGE is a recently created register with information on secondary general education in the Netherlands. The register is derived from the secondary general education part of the Base Register Education Numbers maintained by the Information Management Group. This register is also under development. The 1FigSGE is a harmonized register created by the joint effort of the Ministry of Education, Culture, and Science, the Education Inspectorate, the Secondary Education Council, and Statistics Netherlands. Standardized variables and derivation rules are used. Information from school year 2002/'03 is available on pupils in publicly financed secondary general education. This is a total of 1.3 million pupils (Bakker et al., 2008; Ossen and Daas, 2009). It is used for educational statistics.

3. Results and discussion

3.1 Application

Six administrative data sources were evaluated by means of the checklist. Because our primary interest in this study was the usability of the outcome of the checklist, the checklists were not self-administered but filled out in close cooperation between one (or more) of the authors and several users of the data source. These are key staff members of our office involved in: i) contact with the data source keeper, ii) receipt of the data source, and iii) processing/checking of the data source. The answers of the users and any documentation provided by them were used to respond to the questions included in the checklist. On average around 2 hours were spent to complete a checklist. The end results were reviewed by the authors of this paper and reported back to the users. Any corrections and additional remarks made by the users were included in the final version of the completed checklist.

3.2 Scores obtained

The evaluation results obtained for the six data sources are shown in tables 4 and 5. In table 4 the results for the Source hyperdimension and in table 5 those for the Metadata hyperdimension are shown. For the PA, the Metadata part of the checklist was filled in with its use for the labour statistics in mind. For the other data sources, the envisaged use was educational statistics. Evaluation scores are indicated at the dimension level (compare tables 4 and 5 with tables 1 and 2). Since each dimension contains several quality indicators which are measured by one or more methods, the results shown were obtained by comparing the evaluation results for every measurement method for each quality indicator in each dimension and selecting the most commonly observed score. The symbols for the scores used in table 4 and 5 are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combing symbols with a slash (/) as a separator. An exception is made for unclear results. When in a specific dimension an unclear score occurs for a specific

Table 4: Evaluation results for the Source hyperdimension

DIMENSIONS	DATA SOURCES					
	PA[1]	SFR	CWI	ERR	1FigHE	1FigSGE
1. Supplier	+	+	+	+	+	+
2. Relevance	+	+	+	o	+	+
3. Privacy and security	+	+	+	+	+	+/o
4. Delivery	o	+	-	+	+	o
5. Procedures	+	+/o	+	+/o	+/o	+/o

[1] PA, Policy record Administration; SFR, Student Finance Register; CWI, register of the Centre for Work and Income; ERR, Exam Results Register; 1FigHE, coordinated register for Higher Education; 1FigSGE, coordinated register for Secondary General Education.

Table 5: Evaluation results for the Metadata hyperdimension

DIMENSIONS	DATA SOURCES					
	PA	SFR	CWI	ERR	1FigHE	1FigSGE
1. Clarity	+	+	-	o	+	+
2. Comparability	+/o	+	-	+	+	+
3. Unique keys	+	+	+	+	+	+
4. Data treatment	+/o	?(+)	?	?(o)	?(+)	?(+)

quality indicator this score is shown for the whole dimension. Only when the scores for the other indicators in that dimension are not unclear, the most commonly observed score for those indicators is added between brackets.

3.3 Source discussion

The results in table 4 reveal that the major problem at the Source level is related to the delivery of the CWI. The CWI is hardly ever delivered on time; a delay of a few days or a week is not uncommon. There even has been a period of three months during which no data was delivered at all. Compared to the other data sources, the general score of the 1FigSGE also appears somewhat low. This is, however, not unexpected for a data source in its infancy; it is a relatively new data source. The main problem for the 1FigSGE is delivery related. Because of the recent start of the 1FigSGE, delivery times still fluctuate. On a dimensional level, the scores for all data sources are somewhat low on the delivery and procedures dimension. For the first dimension this is predominantly caused by the not always timely delivery of some of the data sources. This implies a possible risk for the NSI when it relies heavily on the timely availability of these data sources. For the procedures dimension, the scores are somewhat low because of the low scores on the fall-back scenario indicators. Not for all data sources such a scenario has been developed which is not unexpected (Daas and Arends-Tóth, 2009). The scores for this indicator were affected in a negative way because not all information was provided to the users to interpret those questions as they were intended (explained below). With this in mind, hardly any procedural problems were observed. Users can, for example, easily contact the data source keeper in case of trouble and modifications in the data were in most cases clearly and timely communicated.

3.4 Metadata discussion

The results for the Metadata hyperdimension are shown in table 5. Compared to the Source hyperdimension (table 4) more poor (-) scores are observed. Here again the CWI attracts attention. This data source scores negative in the clarity and comparability dimensions. For both dimensions, this is largely the result of the discrepancy between the definition of the CWI-variable ‘level-of-education’ and the definition of the corresponding variable of Statistics Netherlands. A study revealed that the interpretation of the ‘level-of-education’ variable at CWI is highly affected by the combination of the study history and discipline of a job-seeker and the jobs

available (Bakker et al., 2008). For instance, university graduates with a discipline for which almost no jobs are offered at that point in time, are likely to be offered a retraining at a lower level of education to increase their chances for finding a job. When they finish this retraining, the CWI will 'downgrade' their level of education. For some job-seekers, however, it was found that the level of education was upgraded by awarding them the degree of a study they had previously dropped out. CWI clearly has a more practical, less strict, interpretation of the 'level-of-education' variable than Statistics Netherlands. The ERR also scores somewhat low on the clarity dimension because the metadata of the variable definitions for this data source are difficult to interpret.

The data treatment dimension is the most unclear area for nearly all of the data sources. This revealed that in our office hardly any information is available on the checks and modifications of the data performed by the data source keeper. A positive exception in the data treatment dimension scores is the PA. For the PA, specifically in the beginning of its use, Statistics Netherlands has regularly reported problems (at an anonymous level) to the data source keeper that were found to be caused by incorrect working data checks. Although many of the users at Statistics Netherlands are highly interested in the data checks used and the data modifications done by data source keepers it is to be expected that some of the data source keepers, for instance the Dutch Tax administration, are not likely to reveal their checks and modifications in great detail. Despite of this, the NSI should try to gain as much information as possible about the data checks and modifications used. This is certainly a topic that requires more attention.

3.5 Use of the checklist

Apart from the quality related results the users also provided valuable feedback on the usability of the checklist. Based on this feedback some adjustments have been made. One of the major problems was the interpretation of the questions for the indicators in the Unique keys dimension (see Annex). Here, it was not immediately clear to some of the users how these questions should be interpreted. Even when objects are uniquely linked to a key, such as the Citizens Service Number (CSN) for persons in the Netherlands, these keys could still occur more than once in a particular data source. Some users interpreted that as the fact that the CSN-numbers were not unique for the data source. This was not the way the question should have been interpreted. The other major problem was related to questions on the fall-back scenario indicator. The policy of Statistics Netherlands demands that fall-back scenarios need to be developed only for secondary data sources used by the 'image-relevant' statistics (Daas and Arends-Tóth, 2009), which are statistics for which the non-timely publication offers significant risks for the image and the clients of Statistics Netherlands. This essential fact was not included in the question and needs to be added. All feedback provided by the users was used to improve the checklist. In addition the checklist was also reviewed thoroughly by our colleagues from the questionnaire lab.

4. Concluding remarks

The results described in this paper show that the quality framework developed for administrative and other secondary data sources and the corresponding checklist are valuable tools for the evaluation of the statistical usability of those sources. Because the completion of the checklist for the Source and Metadata hyperdimension does not require a lot of time, it is recommended to always start an evaluation of the quality of a secondary data source by filling in the checklist. This should be made a standard procedure for secondary data sources. Advantage of the use of the checklist is that it: i) provides a structured way of looking at the Source and Metadata quality aspects and that ii) not immediately a great deal of attention and work is put into data related quality aspects. The latter is often the case in practice. For the CWI, for example (see tables 4 and 5), attention should first focus on the Source and Metadata level of quality and certainly not on the Data level. When problems regarding the Source and Metadata hyperdimensions cannot be solved satisfactorily, it does not make sense to spend a lot of time and effort in the determination of the quality of the data. For the other data sources evaluated it can be argued that some of the Source and Metadata quality aspects require attention, but overall no serious problems were found. For these sources the quality aspects included in the Data hyperdimension should be determined (Daas et al., 2008b). This hyperdimension is the focus of current research. Main topics being studied are the development of a structured approach for efficiently evaluating the large number of quality indicators in this hyperdimension (table 3) and the use of standardized scripts or software tools to enable a quick determination of those indicators.

Acknowledgments

The authors would like to thank all the ‘users’ at Statistics Netherlands for their assistance and time spend during the completion of the checklists and their valuable feedback. Prof. Bart Bakker is gratefully acknowledged for giving very constructive comments on earlier drafts of this paper. Dr. Marco Puts is gratefully acknowledged for thoroughly reviewing an earlier draft of the checklist and this paper.

References

Bakker, B.F.M. (2009). Micro-integration. *Methodology series* 09001 (in Dutch), Statistics Netherlands, The Hague/Heerlen.

Bakker, B.F.M., Linder, F., Van Roon, D. (2008). Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. *Paper for the International Association for Official Statistics conference on Reshaping Official Statistics*, 14-16 October, Shanghai, China..

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

Cobben, F., Schouten, B. (2008). An empirical validation of R-indicators. *Discussion paper* 08006, Statistics Netherlands, Voorburg.

Daas, P.J.H., Arends-Tóth, J. (2009). Secondary data collection. *Methodology series* 09002 (in Dutch), Statistics Netherlands, Heerlen.

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008a). Proposal for a Quality Framework for the Evaluation of Administrative and Survey Data. *Paper for the workshop on the Combination of surveys and administrative data*, 29-30 May, Vienna, Austria.

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008b). Quality Framework for the Evaluation of Administrative Data. *Paper for the Q2008 European Conference on Quality in Official Statistics*, 9-11 July, Rome, Italy.

Daas, P.J.H., Fonville, T.C. (2007). Quality control of Dutch Administrative Registers: An inventory of quality aspects. *Paper for the seminar on Registers in Statistics - methodology and quality*, 21-23 May, Helsinki, Finland.

Daas, P.J.H., Ossen, S.J.L., Arends-Tóth, J. (2009). Framework of Quality Assurance for Administrative Data Sources. *Paper for the 57th session of the International Statistical Institute*, 16-22 August, Durban, South Africa.

ESC (2007). Pros and cons for using administrative records in statistical bureaus. *Paper for the seminar on increasing the efficiency and productivity of statistical offices*, 11-13 June, Geneva, Switzerland.

Ossen, S.J.L., Daas, P.J.H. (2009). Quality determination of data sources used for the educational register: Source and Metadata hyperdimensions. Internal CBS-paper (in Dutch), DMK-10-04-2009-SOSN, Statistics Netherlands, Heerlen.

Schouten, B., Cobben, F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. *Discussion paper* 07002, Statistics Netherlands, Voorburg.

Statistics Netherlands (2008). *Quality declaration of Statistics Netherlands*. Retrieved from the Statistics Netherlands website, 28 April.

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

Kuijvenhoven, L., Schouten, B. (2008). Quality indicators of the Data hyperdimension. Internal CBS-paper (in Dutch), DMV-2008-03-31-BSTN-LKYN, Statistics Netherlands, Voorburg.

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.

Van Delden, A. and Aelen, F. (2008). Redesigning the chain of economic statistics at Statistics Netherlands: STS-statistics as an example. *Paper for the International Association for Official Statistics conference on Reshaping Official Statistics*, 14-16 October, Shanghai, China.

Van der Laan, P. (2000). Integrating administrative registers and household surveys. *Netherlands Official Statistics*, 15 (2), pp. 7-15.

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

Wallgren, A., Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, UK.

Annex, the checklist

Statistics Netherlands
Division of Methodology and Quality
Methodology Sector Heerlen

Checklist

Quality of Secondary Data Sources

Name secondary data source:

Evaluator(s) name and e-mail address:

Assessment dates:

Start date:

End date:

Checklist for the determination of the quality of secondary data sources

Introduction

Statistics Netherlands and other National Statistical Institutes (NSI's) are increasingly making use of secondary data sources, such as registers, for the production of statistics. An important characteristic of secondary data sources is the fact that they are collected and maintained by other organizations, usually for non-statistical purposes. Since the production of high quality statistics depends on the quality of the input data, it is of vital importance that NSI's have a procedure available to determine, in a systematic, objective, and standardized way the quality of secondary data sources. For this purpose a quality framework was developed that specifically focuses on the quality, i.e. statistical usability, of secondary data sources. The framework consists of three high level views on the quality of a data source; these views are called: Source, Metadata, and Data. Evaluation of the quality indicators in the Source and Metadata views occurs by filling in this checklist.

The checklist is specifically developed to determine the statistical usability of secondary data sources at a general level. Very specific checks are not included because: i) it is impossible to include all possible specific checks, and ii) different users of a data source may have different population parameters in mind that pose different quality constraints. Necessarily, the quality framework has to be restricted to some extent as it is impossible to meet all conceivable uses. If specific checks are required these should only be done after the general evaluation has been performed. Advantage of the more general approach is the fact that it enables the comparison of the quality across time and domains.

Who should use it?

The checklist should be filled in by an internal (future) user of the data source and/or an expert for the secondary data source. For the Source part it is advised to contact the NSI contact person for the particular data source (if available).

Purpose of the Checklist

The checklist consists of two parts. In the first part, the Source view, the quality aspects of the data source as a whole, the data source keeper, and the delivery of the data source are evaluated. In the second Metadata part, the metadata related quality aspects of the data source are determined. Here, also some process related metadata quality aspects are included. Each view is composed of several quality dimensions which each contain a number of quality indicators. Each quality indicator is scored by filling one or more questions in the checklist.

Filling in the checklist starts with the Source part. When problems are found or a quality indicator question cannot be answered completely, the user is guided in the steps to take.

When all Source related questions are answered to completion and the user has a use in mind for the secondary data source, he/she can start the evaluation of the Metadata part of the checklist. In all other cases, evaluation is halted and the data source cannot be used by the NSI.

To increase the efficiency and speed of filling in the checklist, routing instructions (for example: → Go to 1.2 Variable definition) and actions (for example: → Contact the data source keeper) are included.

Examples

1.1.a. Contact 1: data source name	
What is the name of the data source? *Include the internet address if appropriate*	

2.4.a. Response burden 1: Expected consequences	
What is the expected effect of the use of the data source on the response burden of the NSI?	**1**: increase of response burden **2**: no chance → **Go to 3. Privacy and security** **3**: decrease of response burden **0**: don't know → **Go to 3. Privacy and security**

SOURCE

The evaluation of the secondary data source starts with the quality indicators of the Source part listed on the next page. The quality indicators in the Source part are grouped in five dimensions, namely: *Supplier, Relevance, Privacy and security, Delivery,* and *Procedures*.

SOURCE: 1. Supplier

1.1.a. Contact 1: Name of the data source	
What is the data source's name? *Include a reference to internet address if applicable*	

1.1.b. Contact 2: Data source keeper contact information	
Contact information of the organisation that collects and creates the data source. *Even when information is incomplete or lacking, the data that is available must be noted. The fact that data is missing should also be noted.*	Name of the organisation:
	Street name and number: Postal/Zip code and city:
	Name of the contact person:
	Telephone number of the contact person:
	E-mail address of the contact person:
	Function and organisational unit (department) of contact person:
	Other information:

1.1.c. Contact 3: Data source provider contact information	
Record the contact information of the data source provider, if not identical to that of the organisation that collects and creates the data source	Name of the organisation and contact person:
	Telephone number of the contact person:
	E-mail address of the contact person:

1.1.d. Contact 4: NSI contact person information	
Contact information of the NSI contact person for the data source	Name:
	Telephone number:
	Division and department:
	Additional information:

1.2. Purpose: Reason for use	
What is the reason for use of the data source by the data source keeper? *Why is the data source keeper collecting data and maintaining the data source?*	

SOURCE: 2. Relevance

2.1.a. Usefulness 1: Replacement	
The data source is potentially suited **to replace** the data collection process of the following statistics: *List a maximum of 3 statistics and mark to which degree you think the data source will be useful:* **1**: partly useful **2**: useful **3**: very useful **0**: don't know	1. …….. ……**Score: 1 2 3 0** 2. …….. ……**Score: 1 2 3 0** 3. …….. ……**Score: 1 2 3 0**

2.1.b. Usefulness 2: Supplemental use/ check	
The data source is potentially useful to **supplement/check** the following statistics: *List a maximum of 3 statistics and mark to which degree you think the data source will be useful. Statistics already using the data source should also be included here:* **1**: partly useful **2**: useful **3**: very useful **0**: don't know	1. …….. ……**Score: 1 2 3 0** 2. …….. ……**Score: 1 2 3 0** 3. …….. ……**Score: 1 2 3 0**

2.2. Envisaged use	
The data source is potentially useful for the following **new** statistics: *List the name or describe the statistics*	

2.3. Information demand	
How important is the data source for the NSI? *Mark the score (1,2,3, 0) that you find appropriate*	**1**: not that important **2**: important **3**: very important **0**: don't know
Briefly describe the importance of the data source to the NSI.	

2.4.a. Response burden 1: Expected consequences	
What is the expected effect of the use of the data source on the response burden of the NSI?	**1**: increase of response burden **2**: no chance → **Go to 3. Privacy and security** **3**: decrease of response burden **0**: don't know → **Go to 3. Privacy and security**

2.4.b. Response burden 2: Expected effect in hours	
What is the expected effect on the response burden in hours? *Calculation: net effect = (number of questionnaires normally send – number of questionnaires send when data source is used) * average fill in time.* *An estimation may also be given*	

SOURCE: 3. Privacy and security

3.1. Legal provision	
Is there a law, act, or other legal agreement on the basis of which the data source is being maintained? *Include a reference to the law, act or legal agreement*	**1**: no **2**: yes, namely…… …… …… ……..*(briefly describe the legal basis)* **0**: don't know

3.2.a. Confidentiality 1: Data Protection Act	
Does the National Data Protection Act or European Data Protection directive apply to the data in the source?	**1**: no → **Got to 3.3a Security 1** **2**: yes **0**: don't know

3.2.b. Confidentiality 2: Reported use	
Is the use of the data source reported to the organisation that oversees the processing of personal data in accordance with the provisions laid down in the Data Protection Act?	**1**: no → **Contact the organisation (or local representative)** **2**: yes **0**: don't know

3.3.a. Security 1: Data submission	
In what manner will the data be transferred to the NSI? *Describe the manner used, such as: tape, FTP, e-mail, DVD etc.*	

3.3.b. Security 2: Data security arrangements	
Are special arrangements required for the secure submission of data?	**1**: no **2**: yes, namely……. …….. …….. …….*(briefly describe the arrangements)* **0**: don't know

3.3.c. Security 3: Special hardware/software	
Does the NSI have to purchase any special hard- and/or software to enable the secure submission?	**1**: no **2**: yes, namely…… …… …… ……..*(briefly describe hard/software; name, type and brand)* **0**: don't know

SOURCE: 4. Delivery

4.1. Costs	
Are any costs involved in the use of the data source? *Fill in the amount, period, and number of deliveries.*	**1**: no **2**: yes, namely……..…*per day/ month/year for a total of ……. deliveries.* **0**: don't know

4.2.a. Arrangements 1: Terms of delivery	
Are the terms of delivery documented?	**1**: no **2**: yes, in a single general contract **3**: yes, every type of delivery is specified in a separate document **0**: don't know

4.2.b. Arrangements 2: Frequency of delivery	
How often is the data delivered? *Describe the current or expected situation.*	**1**: on request **2**: on regular intervals. ……..……..…… *(report frequency)* **0**: don't know

4.3.a. Punctuality 1: Current delivery	
How punctual is the data delivered? *Describe the current situation.*	**1**: delivery dates/times varies; a delay of ……. is quit common ……..……..*(months, weeks, days, hours)* **2**: delivery is always on time **0**: don't know

4.3.b. Punctuality 2: Delays reported	
When a delay occurs, is this reported in time to the NSI? *Describe the current situation.*	**1**: no, the NSI is not informed **2**: yes, the NSI is informed on time **0**: don't know **9:** not applicable: deliveries are always on time.

4.3.c. Punctuality 3: Data storage	
How quick is new or changed data stored by the data source keeper?	**1**: with a delay of ……..…..…..… *(day, hours, minutes)* **2**: immediately **0**: don't know

4.4. Format	
Data format(s) in which the data can be delivered to the NSI	

4.5.a. Selection 1: Data selection	
What data can be delivered to the NSI? *List units and variables* *(give a brief description for data sources that contain large amounts of variables)*	

4.5.b. Selection 2: Requirements	
Does the selection of data that can be delivered comply with the requirements of the NSI?	**1**: no ……..……..……..……..… *(report the data missing)* **2**: yes **0**: don't know

SOURCE: 5. Procedures

5.1. Data collection	
Is the NSI informed about the way the data is collected by the data source keeper?	**1**: no **2**: yes *(check how recent this information is)* **0**: don't know

5.2.a. Planned changes 1: Familiarity	
Is the NSI informed about any changes to the data source and/or its maintenance?	**1**: no **2**: yes, namely……... …… *(report the plans)* **0**: don't know
***If yes*:** What are the expected consequences to the NSI? (*describe briefly*)	

5.2.b. Planned changes 2: Communication of changes	
Is the NSI informed about the way in which changes are reported by the data source keeper?	**1**: no **2**: yes, namely……... …… *(report how changes are communicated)* **0**: don't know

5.3. Feedback	
Is the NSI allowed to ask questions or contact the data source keeper in case of problems? *Consider both general and data source content related contacts*	**1**: no, because……... …… *(describe reason)* **2**: yes **0**: don't know → **Find out if feedback is allowed and how**
If yes: Are there restraints (technical and regarding the content) concerning the feedback of information?	**1**: no **2**: yes, namely……..…. ……. ……*(describe reason)* **0**: don't know

5.4.a. Fall-back scenario 1: Risk estimation	
Estimate the risk for the NSI that the data source is not delivered on time	**1**: low **2**: average **3**: high, namely…… …… *(describe reason)* **0**: don't know
For score 3 (high): Is a fall-back scenario drawn up? (Note: *this is only required for the image-relevant statistics*)	**1**: no **2**: yes **0**: don't know

5.4.b. Fall-back scenario 2: Arrangements	
What arrangements are made when the data source is not or only partially delivered on time? *Briefly describe the arrangements made (take the stability of the delivery into account)*	

SOURCE: 6. Remarks and Conclusions

6.1. Remarks regarding Supplier and contact
Decisions and actions
When the supplier information is incomplete, the data source keeper must be contacted. → **Go to 6.6**

6.2. Remarks regarding Relevance

6.3. Remarks regarding Privacy and security

6.4. Remarks regarding Delivery

6.5. Remarks regarding Procedures

6.6. Conclusion SOURCE	
Does the data source keeper or provider have to be contacted?	**1**: yes → **Submit a request via the NSI contact person of the data source (if appropriate)** **2**: no → **Continue with METADATA**

METADATA

The Metadata part of the checklist focuses on the meta-aspect of the data in the secondary data source. In addition, it also contains some process related meta-aspects. These aspects specifically focus on the data treatment steps performed by the data source keeper. The Metadata part reviews all the information required to understand and use the data in the data source.
To enable the proper evaluation of the Metadata part of the checklist, the user must be aware of the intended use of the data source. When a data source is going to be used for multiple reasons or by more then one statistics, the Metadata part of the checklist has to be filled in for each particular reason or statistics.

Evaluation of the quality of the secondary data source continues on the next page. The quality indicators in Metadata are grouped in four dimensions, namely: *Clarity, Comparability, Unique keys* and *Data treatment by the data source keeper*. The Metadata part of the checklist is scored in a way similar to the Source part.

METADATA: 1. Clarity

1.1. Population unit definition	
Are the population units defined clearly?	**0**: description missing **1**: description unclear/ambiguous **2**: description clear
Describe the population units as defined by the data source keeper	

1.2. Classification variable definitions	
List the names of a maximum of 10 key classification variables and score the clarity of the definition of those variables by the data source keeper: **0**: description missing **1**: description unclear/ambiguous **2**: description clear	1. ……**Score: 0 1 2** 2. ……**Score: 0 1 2** 3. ……**Score: 0 1 2** 4. ……**Score: 0 1 2** 5. ……**Score: 0 1 2** 6. ……**Score: 0 1 2** 7. ……**Score: 0 1 2** 8. ……**Score: 0 1 2** 9. ……**Score: 0 1 2** 10……**Score: 0 1 2**

1.3. Count variable definitions	
List the names of at most 10 key count variables and score the clarity of the definition of the variables by the data source keeper: **0**: description missing **1**: description unclear/ambiguous **2**: description clear	1. ……**Score: 0 1 2** 2. ……**Score: 0 1 2** 3. ……**Score: 0 1 2** 4. ……**Score: 0 1 2** 5. ……**Score: 0 1 2** 6. ……**Score: 0 1 2** 7. ……**Score: 0 1 2** 8. ……**Score: 0 1 2** 9. ……**Score: 0 1 2** 10……**Score: 0 1 2**

1.4. Time dimension	
Is the period or point in time to which the data refer clearly described by the data source keeper?	**0**: description missing → **Contact the data source keeper** **1**: description unclear/ambiguous **2**: description clear
Describe the time interval of the data in the source as indicated by the data source keeper.	

1.5. Definition changes

When the data source keeper has adjusted a definition, is this change communicated clearly?	**1**: no **2**: yes **0**: don't know **9:** not appropriate: no changes have occurred
If yes (score 2): Which definitions have changed?	

1.5. Decisions and actions
When one or more of the above quality indicators are scored 'description unclear' (score 1) or 'description missing' (score 0) the data source keeper needs to be contacted. Only when these issues are solved, evaluation may continue from here on. In all other cases evaluation stops here.

METADATA: 2. Comparability

Remark: This dimension is not or less relevant for new statistics

2.1. Comparability of the population unit definition	
How comparable are the definitions of the population units used by the data source keeper and the NSI?	**0**: description missing / information absent **1**: unequal, conversion is impossible **2**: unequal, conversion is possible **3**: equal (100% identical)

2.2. Comparability of classification variable definitions	
How comparable are the definitions of the classification variables used by the data source keeper and the NSI? ***Compare the same variables as listed in 1.2.*** *Score the comparability of the variables by marking the appropriate value*: **0**: description missing **1**: unequal, conversion is impossible **2**: unequal, conversion is possible **3**: equal (100% identical)	1. **Score: 0 1 2 3** 2. **Score: 0 1 2 3** 3. **Score: 0 1 2 3** 4. **Score: 0 1 2 3** 5. **Score: 0 1 2 3** 6. **Score: 0 1 2 3** 7. **Score: 0 1 2 3** 8. **Score: 0 1 2 3** 9. **Score: 0 1 2 3** 10. **Score: 0 1 2 3**

2.3. Comparability of count variable definitions	
How comparable are the definitions of the count variables used by the data source keeper and by the NSI? ***Compare the same variables as listed in 1.3.*** *Score the comparability of the variables by marking the appropriate value:* **0**: description missing **1**: unequal, conversion is impossible **2**: unequal, conversion is possible **3**: equal (100% identical)	1. **Score: 0 1 2 3** 2. **Score: 0 1 2 3** 3. **Score: 0 1 2 3** 4. **Score: 0 1 2 3** 5. **Score: 0 1 2 3** 6. **Score: 0 1 2 3** 7. **Score: 0 1 2 3** 8. **Score: 0 1 2 3** 9. **Score: 0 1 2 3** 10. **Score: 0 1 2 3**

2.3. Time differences	
Are the time references used by the data source keeper comparable to those used by the NSI?	**0**: description missing **1**: unequal, conversion is impossible **2**: unequal, conversion is possible **3**: equal (100% identical)

2.4. Decisions and actions

When the data source is used to replace or is used in addition to other data sources and some of the comparability indicators have scored ‘unequal and conversion is impossible’ (score 1) or ‘description missing’ (score 0), the data source cannot be used and the evaluation stops here. These scores are less relevant for Data sources that are used for new statistics. In the latter and all other cases, evaluation may continue.

METADATA: 3. Unique keys

3.1. Presence of unique keys	
Is a unique key present that can be used to identify the population units? *When more unique keys are present list the most appropriate and important one*	**1**: no **2**: yes, namely…… *(name of most important unique key)* **0**: don't know
If yes (score 2): Is the unique key comparable to a unique key used by the NSI?	**0**: description missing **1**: keys unequal, conversion is impossible **2**: keys unequal, conversion is possible **3**: keys equal (100% identical)

3.2. Presence of unique combinations of variables	
Are combinations of variables present that can be used to uniquely identify the population units?	**1**: no **2**: yes, namely…… …… …… …… …… *(list the combinations)* **0**: don't know **9**: not appropriate

3.3. Decisions and actions
Data sources that need to be linked to other sources and were found not to contain unique keys or unique combination of variables, cannot be used. When this is the case, evaluation should stop here. When the presence of unique keys or unique combination of variables is not known for a data sources, this should be investigated in more detail. Contacting the data source keeper might be required to solve this problem. In all other cases evaluation may continue.

METADATA: 4. Data treatment by the data source keeper

4.1.a. Checks 1: Population units	
Does the data source keeper check the population units?	**1**: no **2**: yes, namely…… *(describe check used)* **0**: don't know

4.1.b. Checks 2: Variables	
Does the data source keeper check variables? (e.g. range checks)	**1**: no **2**: yes, namely…… *(describe check used)* **0**: don't know

4.1.c. Checks 3: Combination of variables	
Does the data source keeper check the plausibility of variable combinations? (e.g. edit rules)	**1**: no **2**: yes, namely…… …… …… *(describe checks used)* **0**: don't know

4.1.d. Checks 4: Extreme values	
Does the data source keeper check for the occurrence of extreme values?	**1**: no **2**: yes, namely…… …… …… *(describe checks used)* **0**: don't know

4.2. Modifications	
Does the data source keeper modify (impute, edit, use default values) data?	**1**: no **2**: yes, namely…… …… …… …… …… *(describe modifications used)* **0**: don't know
If yes (score 2): Are modified values marked in the data source and is the original data included or available?	**1**: no **2**: yes **0**: don't know

4.3. Decisions and actions
If in one or more of the above indicators a 'don't know' (score 0) is answered, the data source keeper needs to be contacted to clarify these issues.

METADATA: 5. Conclusions

5.1. Remarks regarding Clarity

5.2. Remarks regarding Comparability

5.3. Remarks regarding Unique keys

5.4. Remarks regarding Data treatment (by data source keeper)

5.5. Conclusion METADATA 1	
Is every question for each indicator answered?	**1**: no, because…… …… …… …… …… *(describe which not and why)* **2**: yes

5.5. Conclusion METADATA 2	
Do all the indicators in the *Clarity*, *Comparability*, and *Unique key* dimensions have a score of 2 or higher and in the *Data treatment* dimension a score of 1 or higher?	**1**: no **2**: yes → **Go to Data-part of the evaluation procedure (*under development*)**
If no (score 1): *Is this a problem for the NSI?*	**1**: no, because…… …… …… → **Go to Data-part of the evaluation procedure (*under development*)** **2**: yes, because…… …… …… …… …… → **STOP EVALUATION**

Room for additional remarks