International Journal for Quality research
UDK- 001.103.2(497.11)
Original Scientific Paper (1.01)

METADATA QUALITY EVALUATION OF SECONDARY DATA SOURCES

Piet J.H. Daas[1]
Saskia J.L. Ossen[1]

[1] *Division of Methodology and Quality, Statistics Netherlands, Room 1C34, CBS-weg 11, 6412 EX, Heerlen, the Netherlands*

Abstract: *Quite some researchers use data sources collected by others for their own investigations. Because of this choice, the research becomes highly dependent on the quality of the secondary data sources used. Therefore, it is important that, prior to use, one is able to check the quality of secondary data sources in a systematic and standardized way. This paper describes the development of a procedure for the evaluation of the quality of the metadata of secondary sources. In it two different views on metadata quality are distinguished. The first view focuses on quality aspects essential for the delivery of the source, whereas the second view predominantly focuses on the conceptual metadata aspects. The quality is determined with a checklist. The approach, originally developed for secondary data sources used for statistics, is applicable to all areas of research that make use of secondary data sources.*

Keywords: *metadata quality, secondary data sources, checklist*

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 (UNECE 2007). Administrative data sources and registers are examples of this (Wallgren and Wallgren 2007). The general term commonly used to identify these sources is secondary data sources (Hox and Boejje 2005). The data in secondary data sources is used in a different way or to answer a different research question then intended by those who originally collected the data. Apart from statistics, the re-use of already existing data is found in quite some other scientific disciplines.

Examples of this are found in the medical, social, and information sciences; for more details see Sørensen et al. (1996), Schutt (2009), and Knight and Burn (2005). Although the work described in this paper was originally developed for statistics the results are applicable to all other areas of research that make use of secondary data sources, specifically when used on a regular basis.

During the last decade, more and more NSI's have started to use secondary data sources (UNECE 2007). A major advantage of these sources is that they drastically reduce the costs of data collection and reduce the administrative burden on enterprises and persons (Statistics Finland 2004). However, secondary data also has some disadvantages. For example, the collection and processing of the data is beyond the control of the NSI. It is the data source keeper who manages these aspects. The same is true for the units and variables a secondary data source contains. These are often defined by organization specific 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, i) an NSI uses a secondary data source for a purpose different than the one for which the data was originally collected and ii) the population over which the NSI wants to publish statistics is not identical to the one over which the data source keeper collects data. As a result of these differences, the quality of the data source, e.g. the 'statistical' usability, needs to be thoroughly studied by an NSI prior to its use.

This often takes considerable effort (Bakker 2009; Bakker et al. 2008; ESC 2007; Van der Laan 2000). Since NSI's want to produce high quality statistics - 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 secondary data sources as early as possible and in an efficient and standardized way. Quality applies both to the data and the metadata domain of the data sources used (Batini and Scannapieco 2006).

Although secondary data has been used by NSI's for quite some time, the determination of the quality of those data sources has not received a lot of attention (UNECE 2007; Saebø et al. 2003; Statistics Finland 2004). Most of the quality studies performed at NSI's have focused on the quality of data collected by surveys (Biemer and Lyberg 2003; Van den Brakel et al. 2007), and on the quality of the statistics produced (Eurostat 2003a-b; 2005b). Only a relative small number of studies have investigated the quality aspects of secondary data used for statistical purposes (Daas et al. 2008).

Moreover the predominant focus of these studies was always on the quality of the data. To our knowledge, metadata related quality aspects of

secondary data sources have never been thoroughly addressed. This is striking, because it is a well-known fact that data is useless without a complete understanding of its accompanying metadata (Struijs 1999).

For data collected and maintained by others this is especially the case. The determination of the quality of the metadata of secondary data sources is the focus of this paper. The paper is structured as following. Section two gives an overview of the quality aspects of secondary data sources. Results of the literature and an object oriented model study are presented.

Metadata related quality aspects are highlighted and a framework for the determination of metadata quality is constructed. Section three deals with the checklist developed for the evaluation of the metadata quality aspects and introduces the eight data sources studied. In section four, the evaluation results of those data sources are presented and discussed.

Section five ends the paper. In this section the results and the significance of the findings for other research areas are discussed. Please note that the word 'aspect' is used in this paper to describe any measurable part of quality.

2. A QUALITY FRAMEWORK FOR METADATA

Quality is a multidimensional concept. This observation applies both to the data and the metadata domain of quality (Batini and Scannapieco 2006). Both domains are important when studying the quality of data sources. For NSI's, quality has become even more important since the adoption of the European Statistics Code of Practice (Eurostat 2005a). As a result of this adoption, NSI's of European Union (EU) member states have committed themselves to an encompassing approach towards high quality statistics. NSI's of the EU-member states involved and NSI's of some other European countries, such as Norway, report the quality of their statistical output by using six quality dimensions.

The dimensions used are: Relevance, Accuracy, Timeliness and punctuality, Accessibility and clarity, Comparability, and Coherence (Eurostat 2005b). Both data and metadata related quality aspects are included in this framework that was specifically developed for the statistical output of NSI's (Eurostat, 2003a).

For the determination of the quality of data sources used for the input of NSI's, such as secondary data sources, the six quality dimensions of Eurostat can not be directly applied (Eurostat 2003b; Daas and Fonville 2007). To what extent these dimensions can be used for the determination of the quality of the metadata of secondary data sources is even more unclear. With these observations, the work described in this paper started.

2.1 Quality aspect identification

An extensive literature study revealed that the views on the composition of the quality of secondary data sources -to be used for statistics- varied greatly. Only a few publications where found that attempted to construct a quality framework for those types of data sources. On the specific study of the quality aspects in the metadata domain hardly any papers were found; most papers entirely or predominantly focused on data related quality aspects. We therefore started by collecting literature that studied the quality of secondary data sources -to be used for statistics- in general. Any metadata related quality aspects would be deduced from those results later on in the process.

The most important developments in the study of the quality aspects of secondary data sources were found to be described in a limited set of papers and books, these are: Wallgren and Wallgren (2007), Daas and Fonville (2007), Eurostat (2003b), Karr et al. (2006), UNECE (2007), Thomas (2005), and ONS (2005). When the results of these studies were compared, a clear difference between the number and types of quality groups or dimensions identified for the statistical quality aspects of secondary data was observed. In our opinion this pointed out the complexity of the problem but also revealed that every researcher or group of researchers had a slightly different view on this topic. The progress in this field would be considerable if these heterogeneous views could be combined somehow into a single framework. This exercise was performed by the authors of this paper. It provided information on both data and metadata related quality aspects. In this paper, the metadata quality related aspects are discussed. Preliminary results for data related quality aspects can be found in the conference papers of Daas et al. (2009a; 2010).By combining the various metadata related quality aspects identified at our office (Daas and Fonville 2007) and those mentioned in the publications of others (listed above), the authors attempted to get a complete overview of all the quality aspects of secondary metadata relevant for statistical use. Some were unique, such as 'Comparability of the area over which data is reported' (Eurostat 2003b) while others were always mentioned, such as 'Clarity of the unit definition'. Whenever possible, every quality aspect included in a study was compared with those observed in any of the other studies. Any overlap between the aspects found was also taken into consideration. The results obtained were discussed and reviewed a number of times. During this exercise also some missing quality aspects were identified, for instance 'Countermeasures when a data source is not delivered on time'; i.e. availability of a fall-back scenario. During this work two important findings emerged. The first one is the fact that -despite of the differences observed- there is a clear general level of mutuality. In a lot of studies many

(very) similar metadata related quality aspects were identified. The second one is the observation that the statistical quality of secondary metadata is more than a multidimensional concept. Depending on the perspective from which the data source is looked upon, different quality aspects prevail. Such a perspective -a high level view at statistical quality- is nothing new, it has been described several times by others.

These views are called categories (Batini and Scannapieco 2006) or hyperdimensions (Karr et al. 2006). The latter term will be used in the remainder of this paper. A hyperdimension is a way of looking at quality at a level higher than that of the commonly used dimension.

In a hyperdimension several dimensions of (metadata) quality are grouped. The quality aspects included are highly influenced by the contextual view on the quality of the data source (Karr et al. 2006). With the above in mind, a quality framework was developed for secondary metadata that consisted of hyperdimensions, dimensions, quality indicators, and measurement methods (figure 1). A hyperdimension is composed of two or more dimensions and each dimension contains one or more quality indicators. A quality indicator is measured or estimated by one or a combination of methods. The relation between the various quality aspects included in the framework is shown in figure 1.

To check the completeness of the quality aspects identified another study was additionally performed. Here, quality aspects were identified by means of the Object Oriented Quality Management (OQM) model (Van Nederpelt 2009a-b). The OQM-model is compatible with the well known European Foundation for Quality Management Excellence model (EFQM) but has the additional advantage that it conforms to the European Statistics Code of Practice (Eurostat 2005a) and the Quality Declaration of the European Statistical System (Eurostat 2002). Both the Code and the Declaration are obviously very important for NSI's of EU-member states.

The general purpose of the OQM-approach is to identify areas of quality for an object that an NSI wants to control (Van Nederpelt 2009a). The object used in the OQM-exercise was 'secondary data source'.

The outcome of the application of the OQM-model was a list of quality dimensions and aspects specific for secondary data sources (Daas and Van Nederpelt 2010). Comparison of this list with those identified by the literature study revealed that no quality aspects were missing from the latter list.

The outcome fully supported the completeness of the list of quality aspects found for secondary metadata.For more information on OQM and for a more detailed description of the results of OQM-approach the reader is referred to the papers of Van Nederpelt (2009a-b) and Daas and Van Nederpelt (2010), respectively.

2.2 Metadata quality aspects

The identification and comparison of all the quality aspects identified for secondary data sources, i.e. both data and metadata related, revealed three discernible contextual ways of looking at the quality of such a data source.

The three hyperdimensions identified were called: Source, Metadata, and Data. The quality indicators in these hyperdimensions all have a product based approach in common (Ehling and Körner 2007). The quality indicators in the three hyperdimensions do not overlap; they each highlight different quality aspects of the data source (Daas et al. 2009a).

The combined set of indicators for the Source and the Metadata hyperdimension contain all quality indicators specific to the metadata domain. The quality indicators for the data domain are all included in the Data hyperdimension. For more information on the latter hyperdimension the reader is referred to the conference papers of Daas et al. (2008; 2010).

Although the quality indicators in the Source and Metadata hyperdimension both describe quality aspects that belong to the metadata domain of quality, their content differs considerably. The quality indicators in the Source hyperdimension are predominantly related to the delivery of the source and the access to source.

These are all aspects that need to be resolved before an NSI is able to use the data source for purposes of statistics production on a regular basis. In table 1 the dimensions, quality indicators, and measurement methods included in the Source hyperdimension are listed. Apart from the delivery and access related quality aspects, the Source hyperdimension additionally contains a few aspects that focus on the effect of the use of the data source on the NSI. In the Source hyperdimension mainly qualitative methods are present. An exception is the calculation of the effect of the use of the data source on the administrative burden of the NSI (table 1, indicator 2.4). The indicators in the Metadata hyperdimension focus on the conceptual and process related quality aspects of the metadata of the source. In table 2 the dimensions, quality indicators, and measurement methods are listed for the Metadata hyperdimension. The majority of the quality aspects in this hyperdimension are on conceptual metadata oriented. The conceptual aspects check if the NSI fully comprehends the definitions of the units, variables and reporting period(s) of the data source keeper and compares them with those used by the NSI.

The presence of variables that can be used to uniquely identify the population units in the data source (i.e. keys) are also examined. The 'Data treatment by data source keeper' dimension is the only dimension in the Metadata hyperdimension that does not contain conceptual related quality indicators (table 2, dimension 4).

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

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 NSI 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

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

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

It focuses on checks and modifications of the data and the uses of default values by the data source keeper. This is very important process-related meta-information because it highly affects the quality of the secondary data source. The Metadata hyperdimension solely contains qualitative methods.

3. CHECKLIST AND DATA SOURCES

To assist the evaluation of the quality aspects in the Source and Metadata hyperdimension (tables 1 and 2) a checklist was developed (Daas et al. 2009b). It is composed of questions that correspond to the measurement method(s) of every quality aspect included in both hyperdimensions. Apart from this, the checklist also guides the user through the measurement methods. To test the usability of the framework and the checklist, eight data sources were evaluated. The checklist and data sources are described in more detail below.

3.1 The checklist

By answering the questions in the checklist, 'measurements' are taken for every measurement method for the quality indicators included in the Source and Metadata hyperdimensions. Since the majority of the methods in those hyperdimensions are qualitative, usually a score has to be filled in. When problems are found or a question can not be answered completely, the checklist guides the user in the steps to take. Apart from this, additional space is included to write down remarks. The checklist can be used both for assessing the quality of a data source that is already being used and for evaluating a (new) data source that is potentially interesting for producing statistics. For the evaluation of the Metadata hyperdimension it is required that the user has a particular statistical use for the data source in mind. This is due to the fact that the conceptual metadata definitions of units and variables in the data source need to be compared with those of the NSI (Daas et al. 2008). In total the checklist is 18 pages long; because of this it was not included in this paper. The reader can obtain the checklist and its explanation by downloading the paper of Daas et al. (2009b).

To test the usability of the checklist and its usefulness for statistics, eight administrative data sources were evaluated. The data sources are described in the next section. Because our primary interest in this study was the usability of the outcome of the checklist, the checklists were not self-administered but filled in in close cooperation between one (or more) of the authors and several 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. This heterogeneous group of people is referred to as 'the users' in the remainder of this paper. Results of the combined efforts of the users and any documentation provided were used to respond to the questions included in the checklist. On average around two hours were spent to complete a checklist. The end results were reviewed by the authors and reported back to the users. Any corrections and additional remarks made by the users were included in the final version of the checklist. The results in the final version of the checklist are discussed in this paper.

3.2 Data sources

A total of eight secondary data sources of SN were evaluated by means of the checklist. The data sources studied were: Insurance Policy record Administration (IPA), Student Finance Register (SFR), register of the Centre for Work and Income (CWI), Exam Results Register (ERR), the coordinated register for Higher Education (1FigHE), the coordinated register for Secondary General Education (1FigSGE), the National Car Pass register (NCP), and the Dutch Municipal Base Administration (MBA). Each of the data sources is briefly described below.

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

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 an important data source for SN. It is, among other things, used in educational and income statistics.

The CWI-register contains information on job seekers in the Netherlands. The register is maintained by the Institute for Employee Benefit Schemes. It contains information on the (previous) jobs, education, and courses job seekers have taken. Information is supplied to SN 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 by SN 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 registers on secondary education (see the 1FigSGE below)

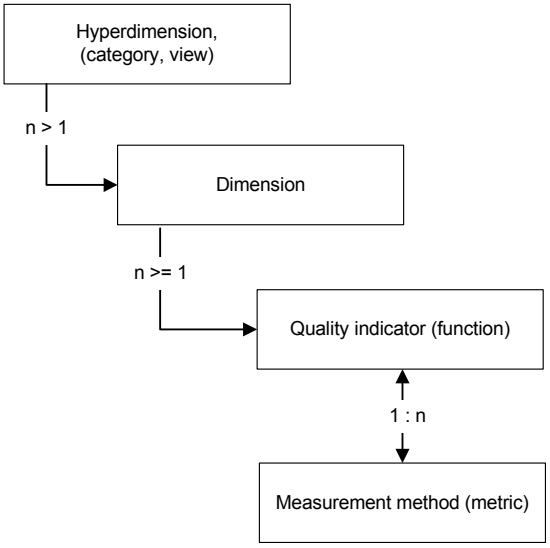

Fig. 1 Hierarchical relation between the different aspects of quality used in the framework developed

The 1FigHE is a register with information on students in 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 SN. 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). The source is used for educational statistics.

The 1FigSGE is a fairly recently created register with information on students in 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. The latter 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 SN. 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). It is used for educational statistics.

The NCP-register is maintained by the National Car Pass Foundation. In the NCP-register the license plate numbers and odometer values of cars in the Netherlands are stored. The odometer values are recorded whenever a car is serviced in a participating garage, car dealer, repair shop, or service station in the Netherlands. A total of more than 6 million odometer values are registered per year.

Main purpose of the NCP-register is the prevention of odometer fraud; i.e. decreasing the number shown on the odometer. For SN, the NCP provides valuable information on the distances driven by cars. The NCP differs from the other data sources because costs are involved for obtaining the data. SN has to pay for the number of records it wants to use. Because of this, each year a 10% part of information of the registered NCP-population is purchased by SN; the samples differ per year. The NCP-register is an important data source for the traffic and transport statistics of SN.

The MBA is maintained by the Dutch Municipalities under the guidance of the Personal Records and Travel Documents Agency, a section of the Ministry of the Interior and Kingdom Relations. The MBA is a decentralized population registration system. This means that every municipality in the Netherlands has its own population register in which personal information of all the inhabitants of that municipality is stored. Examples of the information included in the MBA are: sex, date and place of birth, place of usual residence, marriage date, etc.

Each inhabitant of the Netherlands has a unique identification number. The MBA is an extremely important data source for SN because it is the backbone for all social statistics. SN therefore receives daily updates of all alterations in the MBA (Prins and Kuijper 2007). The MBA is the first base register in the Netherlands.

4. RESULTS

The evaluation results obtained for the eight secondary data sources are shown in tables 3 and 4. In table 3 the results for the Source hyperdimension and in table 4 those for the Metadata hyperdimension are shown.

Table 3. Evaluation results for the Source hyperdimension

Dimensions	*Data sources*							
	IPA	SFR	CWI	ERR	1FigHE	1FigSGE	NCP	MBA
1. Supplier	+	+	+	+	+	+	+	+
2. Relevance	+	+	+	o	+	+	+	+
3. Privacy and Security	+	+	+	+	+	+/o	+	+
4. Delivery	o	+	-	+	+	o	+	+
5. Procedures	+	+/o	+	+/o	+/o	+/o	o	+

For the IPA, the Metadata hyperdimension part of the checklist was filled in with its use for the labour statistics in mind. The MBA was reviewed as a source for the population statistics and the NCP was evaluated with its use for the traffic and trade statistics in mind. The envisaged use of the other data sources was educational statistics. Evaluation scores are indicated at the dimension level (compare tables 3 and 4 with tables 1 and 2).

Table 4. Evaluation results for the Metadata hyperdimension

Dimensions	*Data sources*							
	IPA	SFR	CWI	ERR	1FigHE	1FigSGE	NCP	MBA
1. Clarity	+	+	-	o	+	+	+	+
2. Comparability	+/o	+	-	+	+	+	+	+
3. Unique keys	+	+	+	+	+	+	+	+
4. Data treatment	+/o	?(+)	?	?(o)	?(+)	?(+)	+	+

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 of each quality indicator in each dimension and selecting the most commonly observed score. The symbols for the scores used in table 3 and 4 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 dimension an unclear score occurs for a specific 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 these indicators is additionally shown between brackets.

4.1 Source hyperdimension

The results in table 3 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 weeks is not uncommon. There even has been a period of three months when 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 has relatively recently been created. The main problem for the 1FigSGE is delivery related. Because of the recent start of the 1FigSGE, delivery times still fluctuate somewhat. On a dimensional level, the scores for the data sources are a bit low on the delivery and procedures dimension. For the delivery 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 the data source. For the procedures dimension, the scores are somewhat low because the majority of the data sources scored low on the fall-back scenario indicators. Not for all data sources such a scenario (Daas and Arends-Tóth 2009) has been developed. The scores for the fall-back indicators were also affected in a negative way because not all information was provided to the users to interpret those questions as they were intended; discussed in more detail in section 4.3. With this in mind, hardly any procedural problems were observed, except for the NCP. Here the procedure score is somewhat lower because the data source keeper does not always reply timely. For the other data sources no such problems occurred. Users can easily contact the data source keeper in case of trouble and, in general, the requested information is clearly and timely communicated.

4.2 Metadata hyperdimension

The results for the Metadata hyperdimension are shown in table 4. Compared to the Source hyperdimension (table 3) clearly 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 SN. Detailed 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 SN. SN in contrast, strictly differentiates between the 'true' highest level of education attained (the level of a completed study) and the highest level of education of an attended, but unfinished, study, see Bakker et al. (2006; 2008) for more details. Apart from the CWI, 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.

Of all the dimensions in the Metadata hyperdimension, the data treatment dimension is the most unclear area for the majority of the data sources. This revealed that in our office relatively little knowledge is available on the possible checks and modifications of the data performed by the data source keeper. Positive exceptions are the IPA, NCP, and MBA; here quite a bit is known. For the IPA, especially in the beginning of its use, SN has regularly reported problems (at an anonymous level) to the data source keeper that were eventually found to be caused by incorrect working data checks. For the NCP and MBA detailed information was provided without hesitation by the data source keeper. These are however positive exceptions. Although many of the users at SN are highly interested in the data checks used and the modifications done by data source keepers it is our experience that quit a lot of data source keepers, as for example 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. The IPA, NCP, and MBA demonstrate that it is possible. It is certainly a topic that requires more attention.

4.3 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 small, most textual, and a few major adjustments have been made. One of the major problems was the interpretation of the questions for the indicators in the unique keys dimension (table 2, dimension 3). Here, it was not immediately clear to some of the users how these questions should be interpreted. Even when objects were 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 data source. Some users interpreted that as the fact that the CSN-numbers were not unique for the data source. The latter was not the way the question should have been interpreted. The other major problem was related to questions on the fall-back scenario indicator. Currently the policy of SN 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); the latter are statistics for which the non-timely publication offers significant risks for the image and the clients of SN. This essential fact was not included in the question and needed to be added. All feedback provided by the users was used to improve the checklist. In addition, the checklist was reviewed meticulously by SN-colleagues specialized in developing and testing questionnaires. In the paper of Daas et al. (2009b) the improved version of the checklist is included.

5. CONCLUDING REMARKS

The results described in this paper demonstrate that the quality framework developed for the determination of the metadata of secondary data sources and the corresponding checklist are valuable tools for the evaluation of the statistical usability of such data sources. It is essential that, for every data source, all quality aspects in the Source hyperdimension and all aspects in the Metadata hyperdimension are evaluated. Focus of the latter hyperdimension should, obviously, be limited to the units and variables in the data sources that the user plans to employ. The Metadata hyperdimension also needs to be evaluated with a specific use in mind. When a user plans to use a secondary data source for more than one statistical purpose, the checklist needs to be filled in for each of those purposes. 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 the evaluation of the quality of a secondary data source by filling in the checklist. We highly recommend that this metadata evaluation step is made a standard procedure for all secondary data sources used by an NSI. Advantage of the use of the checklist are that it i) provides a structured way of looking at the quality aspects in the metadata domain and that ii) not immediately a great deal of attention and work is put into the evaluation of data related quality aspects. The latter is often the case in practice.

The evaluation results for the eight data sources described in this paper reveal that attention should be paid to the metadata quality aspects of the CWI before any studies are performed that relate to the data in this source. Only when the problems in the metadata domain

of the CWI are solved satisfactorily, it makes sense to spend (a lot more) time and effort in the determination of the quality of the data of this source. The MBA demonstrates that it is possible to have everything in the metadata domain under control. For the other data sources it can be argued that the results suggest that some of the quality aspects in the Source and/or Metadata hyperdimension require attention. But overall no serious problems were found. For all data sources, with exception of the CWI, the next step would include the study of the quality aspects of the data domain (Daas et al. 2008). The latter domain is also the focus of current research. Main topics being studied are the development of a structured approach for efficiently evaluating the quite large number of quality indicators in this hyperdimension (Daas et al. 2008; 2009a) and the use of standardized scripts or software tools to enable a quick determination of these indicators.

Apart from official statistics, there are many other fields of science that highly rely on secondary data sources as important sources of information. Examples are sociology, criminology, epidemiology, bioinformatics, and medicine; to name a few. For each of these fields a thorough understanding of the quality of the metadata, e.g. the reliability of the delivery of the data source, privacy and security issues, and a clear understanding of the conceptual metadata, is of vital importance. The results described in this paper demonstrate that the metadata framework and the checklist developed can be used for this purpose. The advantages of the use of the checklist is that it assures that the researcher has paid sufficient attention to the set of very important preconditions before he or she start using the secondary data source. Since metadata evaluation only costs a limited amount of time and prevents that some very essential quality aspects are overlooked, we recommend that metadata quality evaluation becomes an essential first step in the use of secondary data sources by any scientific discipline.

REFERENCES:

[1] Bakker, B.F.M. (2009). Micro-integration (in Dutch), Statistical methods 09001, Statistics Netherlands.

[2] Bakker, B.F.M., Bouman, A.M. & Van Toor, L. (2006). Level of education from registers: new data sources but not yet complete (in Dutch). In Engberts, L., Linder, F. & Bastiaans, F. (Eds.), A picture of Social Cohesion, the SSD now and in the future (pp. 141-162), Voorburg: Statistics Netherlands.

[3] 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. Proceedings of IAOS Conference on Reshaping Official Statistics, Shanghai.

[4] Batini, C. & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Berlin: Springer.

[5] Biemer P.P. & Lyberg L.E. (2003). Introduction to Survey Quality. New Jersey: Wiley.

[6] Daas, P.J.H. & Arends-Tóth, J. (2009). Secondary data collection (in Dutch). Statistical methods 09002, Statistics Netherlands.

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

[8] Daas P.J.H. & Fonville T.C. (2007). Quality control of Dutch Administrative Registers: An inventory of quality aspects. Proceedings of the seminar on Registers in Statistics - methodology and quality, Helsinki.

[9] Daas, P.J.H., Ossen, S.J.L. & Arends-Tóth, J. (2009a). Framework of Quality Assurance for Administrative Data Sources. Proceedings of the 57th session of the International Statistical Institute, Durban.

[10] Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R. & Arends-Tóth, J. (2009b). Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands.

[11] Daas, P.J.H., Ossen, S.J.L. & Tennekes, M. (2010). The determination of administrative data quality: recent results and new developments. Paper for the European Conference on Quality in Official Statistics 2010, Helsinki, Finland

[12] 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.

[13] Ehling, M. & Körner, T. (2007). Handbook on Data Quality Assessment Methods and Tools, Wiesbaden.

[14] ESC (2007). Pros and cons for using administrative records in statistical bureaus. Proceedings of the seminar on increasing the efficiency and productivity of statistical offices, Genova.

[15] Eurostat (2002). Quality Declaration of the European Statistical System. Eurostat publication, Luxembourg.

[16] Eurostat (2003a). Definition of quality in statistics, Assessment of the quality in statistics, Item 4.2.

Eurostat publication, Luxembourg.
[17] Eurostat (2003b). Quality assessments of administrative data for statistical purposes, Assessment of quality in statistics, Item 6. Eurostat publication, Luxembourg.
[18] Eurostat (2005a). European Statistics Code of Practice for the national and community statistical authorities. Eurostat publication, Luxembourg.
[19] Eurostat (2005b). Standard quality indicators, Quality in statistics. Eurostat publication, Luxembourg.
[20] Karr, A.F., Sanil, A.P. & Banks, D.L. (2006). Data quality: A statistical perspective. Statistical Methodology, 3, 137-173.
[21] Hox, J.J. & Boeije, H.R. (2005) Data collection, Primary vs. Secondary. Encyclopaedia of Social Measurement Vol. 1, 593-599.
[22] Knight, S-A & Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 8, 159-172.
[23] ONS (2005). Guidelines for measuring statistical quality, version 3.0, Office of National Statistics, London.
[24] Prins, C.J.M. & Kuijper, H. (2007). Population statistics under the person's card system and the MBA-system: similarities and differences (in Dutch). Bevolkingstrends, 55, 14-33.
[25] Saebø H. V., Byfuglien J. & Johannesen R. (2003). Quality Issues at Statistics Norway. Journal of Official Statistics, 19, 287-303.
[26] Schutt, R.K. (2009) Investigating the Social World: The Process and Practice of Research, 5th edition. Thousand Oaks: Pine Forge Press.
[27] 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), 435-442.
[28] Statistics Finland (2004). Use of Register and Administrative Data Sources for Statistical Purposes, Handbook 45, Helsinki.
[29] Struijs, P. (1999). Metadata at Statistics Netherlands. Proceedings of the Federal Committee on Statistical Methodology Research Conference, Arlington, USA.
[30] Thomas, M. (2005). Assessing Quality of Administrative Data. Survey Methodological Bulletin, 56, 74-84.
[31] UNECE (2007). Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics. Geneva: United Nations Publication.
[32] Van den Brakel J., Smith P. & Compton S. (2007). Quality procedures for survey transitions, experiments and discontinuities. Discussion paper 07005, Statistics Netherlands.
[33] Van der Laan, P. (2000). Integrating administrative registers and household surveys. Netherlands Official Statistics, 15, 7-15.
[34] Van Nederpelt, P.W.M. (2009a). The creation and application of a new quality management model. Discussion paper 09040, Statistics Netherlands.
[35] Van Nederpelt, P.W.M. (2009b). The creation and application of a new quality management model. Statistika, 5, 385-395.
[36] Vujović, A., Krivokapić, Z., & Jovanović, J. (2010). Top prioriti QMS principles for achieving business excellence. *International Journal for Quality Research*, 4(2), 117-124.
[37] Wallgren, A. & Wallgren, B. (2007). Register-based Statistics: Administrative Data for Statistical Purposes. England: John Wiley & Sons.

Received: 10.02.2011 Accepted: 10.03.2011 Open for discussion: 1 Year