Quality checklist for registers applied to online price information and offline route information

[*]Saskia J.L. Ossen, Piet J.H. Daas, and Marco Puts
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11, 6412 EX
Heerlen, The Netherlands
sjl.ossen@cbs.nl

Keywords: Register quality, quality checklist, internet prices, route information

1. Introduction

National Statistical Institutes (NSI's) are increasingly using data collected by others for producing statistics. This has the disadvantage that the collection and maintenance of the data used is beyond the control of the NSI. Statistics Netherlands therefore developed a quality framework to determine, in a systematic and standardized way, the quality of registers. This framework distinguishes three hyperdimensions (a way of looking at quality at a level higher than that of the commonly used dimension): Source, Metadata and Data. For the first two hyperdimensions, primarily focusing on the delivery of the data by the data source keeper and the conceptual and process related metadata, a checklist has been developed. This checklist enables a systematic and standardized assessment of the quality aspects belonging to these hyperdimensions. By applying the checklist related to the Source and Metadata hyperdimensions to several registers it has been shown that the checklist is a useful tool for identifying quality related problems in registers (Daas et al., 2009). Current research aims at developing standardized methods for evaluating the quality aspects related to the Data hyperdimension (Daas et al., 2010).

Although registers are an important secondary data source for producing statistics other types of externally collected data sources are also used by Statistics Netherlands, such as: (1) product prices on the internet, and, (2) offline route information for producing road statistics. As Statistics Netherlands currently explores new ways of gathering data (Roos et al., 2009), it is expected that in the future even more external data (of different types) will be used.

Because of these developments it is important to have a quality framework making it possible to determine the quality of these external data sources in a quick, standardized, and objective way. The aim of this paper is therefore to establish whether the checklist developed for registers is also suited for other types of externally collected data. This would firstly have the advantage that there is no need for developing a new checklist for other types of external data. Meaning that it would already be possible to evaluate the Source and Metadata hyperdimensions of these data sources. Another, maybe even more important, advantage would be that applying the same checklist to different types of external data would lead to comparable quality results.

To test the usability of the checklist for other types of external data sources we applied the quality checklist to offline route information and internet prices and we established whether differences do exist between the typical quality problems of registers, offline route information, and prices on the internet. The results are presented in this paper.

The paper is structured as follows. In section 2 the quality framework for registers will be introduced. The introduction of this framework provides useful insights into the quality indicators considered to be useful for registers. As this paper concentrates on the usefulness of the checklist related to the Source and Metadata hyperdimension, the introduction of the quality framework will focus on these hyperdimensions. The offline route information and internet prices to which we applied the checklist are described in section 3. The method for determining the usefulness of the checklist for these data sources will be detailed in section 4. In section 5 the results of applying the

[*] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

checklist are discussed. This means that we will show and discuss the scores of offline route information and internet prices on the quality indicators developed for registers. In section 6 we use these scores and the experiences obtained in applying the checklist to draw conclusions about the usefulness of the checklist for other data sources than registers. We will more specifically consider for every quality aspect included in the checklist whether it also applies to the considered data sources. We also establish whether quality aspects are missing. In the last section overall conclusions are drawn and future research is discussed.

2. The quality framework for registers

The quality framework for registers is composed of several hyperdimensions (Karr et al., 2006), which highlight different aspects influencing the usability of the source for the NSI (Daas et al., 2009).

Three hyperdimensions, viz. Source, Metadata, and Data, are used to determine the statistical usability of the register. Every hyperdimension is composed of several dimensions each containing a number of quality indicators. A quality indicator is measured by one or more measurement

Table 1 Dimensions, quality indicators, and methods for 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 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.

methods (Daas et al., 2008, Daas et al., 2009). The Source, Metadata, and Data hyperdimension each highlight different quality aspects at an increasing level of detail.

In the Source hyperdimension, quality aspects of the data source as a whole, the data source maintainer and the delivery of the data source to the NSI are studied. In Table 1 the dimensions, quality indicators, and measurement methods for the Source hyperdimension are listed. The Metadata hyperdimension specifically focuses on the metadata related aspects of the data source. In Table 2 the dimensions, quality indicators, and measurement methods are listed for the Metadata hyperdimension. The Data hyperdimension focuses on the quality aspects of the data in the data source. These are predominantly accuracy related with the exception of some technical checks (Daas et al., 2009). As the work described in this paper focuses on the quality aspects included in the Source and Metadata hyperdimension, no detailed description of the quality aspects of the Data hyperdimension is given. For details we refer to (Daas et al., 2009, Daas et al., 2010).

For the evaluation of the quality aspects included in the Source and Metadata hyperdimension, a checklist has been developed (Daas et al., 2009). The identical checklist is used in this paper. The checklist guides the user through the quality indicators listed in Table 1 and Table 2. For every measurement method, a question needs to be answered. When problems are found or a question can not be answered completely, the user is also guided in the steps to take.

The quality framework can be applied to (i) registers that are already in use in order to establish whether there are any quality related weak points and to (ii) registers that are not yet used. When the quality framework is applied for the latter case it can be a very useful tool in making well-considered decisions about using the register for producing statistics. The results can also be useful input for negotiations with the data source keeper.

Table 2 Dimensions, quality indicators, and methods for 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

3. Description of data sources examined

In this paper the quality checklist for registers will be applied to two types of external data: i) Offline route information and ii) Product prices on the internet. Both types of external data and their use at Statistics Netherlands (when applicable) are described in sections 3.1 and 3.2.

3.1. Offline route information

Statistics Netherlands produces a statistic related to road traffic. The aim of this statistic is to provide (among others) the number of kilometres driven by Dutch transporters in the Netherlands as well as in foreign countries. To determine the number of kilometres driven in the different countries, Statistics Netherlands uses questionnaires in which a sample of transporters answers questions about the routes they have driven and the corresponding route lengths during a week.

To check and correct the answers given by the transporters, and to impute missing values Statistics Netherlands uses an offline route calculator. In illustration, when a transporter indicates in the questionnaire that he has driven between two places, Statistics Netherlands uses a route calculator to determine whether the corresponding distance is plausible. Also border crossing locations are determined to establish which part of the route has been driven in which country.

As the route calculator is developed by an external company, it can be seen as an external data source for Statistics Netherlands. Statistics Netherlands does, for example, not collect distance information itself but directly uses the information provided by the route calculator. For this sake the quality checklist was applied to the offline route information of several candidate suppliers of the data. The corresponding results were used in the negotiations with the different data source keepers.

3.2. Product prices on the internet

The internet contains a huge amount of information that can provide new data gathering opportunities for Statistics Netherlands. The internet is currently already used by Statistics Netherlands to monitor price developments of products for producing the Consumer Price Index (CPI) (Hoekstra et al., 2010). It is well imaginable that the use of the internet as a source of 'statistical' information will further increase in the future.

The drawback of using information available on the internet is that Statistics Netherlands has to rely on information provided by external parties having other goals than producing statistics. For this reason we also applied the quality checklist. The following price information was checked:

- *Supermarket prices* We considered the website of a large Dutch supermarket chain that enables customers to buy their products online. The prices of the products of this supermarket chain are used for producing the Consumer Price Index. In the latter context Statistics Netherlands does however not collect the information via the internet as it can obtain the required information in a more efficient way (scanner data).
- *Prices of houses* We considered the largest website in the Netherlands showing asking prices of houses that are for sale.
- *Prices of filling stations* We considered the website of a particular unmanned filling station operating throughout the Netherlands.
- *Prices of flight tickets* These prices are already manually collected from several websites to produce the Consumer Price Index.

The websites with flight tickets are selected because Statistics Netherlands uses (comparable) data already. The websites of filling stations and houses are considered as these sites are also examined in a study exploring the use of automated data collection from websites (Hoekstra et al., 2010). As a good test of the checklist requires application to different sites we also added the site with supermarket prices to our investigation (although Statistics Netherlands gets information in a more efficient way). Note that the aim of considering these websites is to establish whether the quality checklist can be applied to internet prices. In this paper we do not use the results of applying the checklist to decide whether the data are of high enough quality to be used by Statistics Netherlands. Based on the aforementioned study on methods for collecting internet data automatically we distinguish two ways of collecting price information: (1) manually, and (2) via software especially developed for automatically collecting prices from a particular website.

4. Methodological approach

To test the usability of the quality checklist for offline route information and internet prices, we applied the checklist to the data sources. To answer the questions related to the offline route calculator we used information provided by the external company and experiences obtained while testing the route calculator. To fill out the checklists regarding the online prices, we visited the aforementioned websites and studied them carefully. Especially the "general conditions" and the "frequently asked questions" webpage's often provided useful information.

The results obtained by filling out the checklists were used in two ways for determining whether the quality checklist for registers can be applied to other types of external data as well:

- We used the scores on the checklist to see whether a particular data source had an extreme score in one of the quality dimensions. The reasoning behind this is that such an extreme score might point at a difference between quality aspects important for registers and quality aspects important for other types of data sources. We also considered the ease of determining a score to establish whether a specific dimension is applicable to a given data source type.
- Next we considered if the evaluation of the data sources pointed out any quality related issues not included in the checklist.

5. Results of applying the quality framework

In this section the scores of the different data sources will be presented. More specifically a summary of the scores regarding the Source hyperdimension is presented in Table 3 and a summary of the scores related to the Metadata hyperdimension is shown in Table 4. In sections 5.1 and 5.2 we respectively discuss the results for the offline route information and the online internet prices in more detail.

Table 3 Evaluation results for the Source hyperdimension

	Offline route information	Internet Prices			
		Supermarket prices	Prices of houses	Prices of filling stations	Prices of flight tickets
Supplier	+	?	?	?	?
Relevance	+	?	?	?	+
Privacy and security	+	+	+	+	+
Delivery	+	+	+	+	+
Procedures	+/ o	o/+	o/+	o	o

Table 4 Evaluation results for the Metadata hyperdimension

	Offline route information	Internet Prices			
		Supermarket prices	Prices of houses	Prices of filling stations	Prices of flight tickets
Clarity	+	+/o	+/o	+/ o	+/ o
Comparability	+	?	?	?	+
Unique keys	+	+	+	+	+
Data treatment	o	+	+	+	+

Note that the evaluation scores are indicated at the dimension level (compare Table 1 and Table 2 with Table 3 and Table 4). 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 3 and Table 4 are: good (+), reasonable (o), poor (-) and unclear (?); intermediary scores are created by combing

symbols with a slash (/) as a separator. In the following sections the scores for the different data sources summarized in Table 3 and Table 4 are discussed in more detail.

5.1. Scores for offline route information

The *supplier* of the offline route information is a commercial company having a delivery agreement with Statistics Netherlands. All contact information is known. This external data deliverer has, at first glance, the same purpose as Statistics Netherlands: determining routes and their lengths. An important difference is however that commercial route calculators spend a lot of effort in using real time information in planning routes. For example, by using information about (recurrent) congestion, route calculators can adjust their route advices in order to avoid congestion. Consequently, advised routes can differ between different times of day. This is a disadvantage for Statistics Netherlands, as it uses the planner to check routes driven in the past for which no information about the time at which the route was driven is available. In fact it is important that entering two locations always leads to the same route (having the same border crossings and so on) and the same distance to make the procedures reproducible. An important feature of the route calculator for Statistics Netherlands is therefore the possibility to switch off the use of real time information. Since this possibility is available the "*relevance* dimension" is scored "+" in Table 3.

Given that Statistics Netherlands uses an offline route calculator there are no *privacy and security* problems (Table 3). The information that is provided to Statistics Netherlands is in fact a roadmap and software for calculating routes, no personal information entered by the transporters is exchanged. It could be different when Statistics Netherlands would use an *on*line route calculator as this would require that (private) information provided by the transporters would be sent to the data source keeper in order to determine the required information.

The *delivery of the data* needs in this context to be interpreted as the delivery of an update of the map used by the route calculation tool and an update of the software. The "deliveries" do not occur at fixed moments in time, but only when significant changes are made. There is at least one delivery each year.

The motivation for scoring the *procedures* with good to reasonable is that Statistics Netherlands does not know in all detail how the algorithm used by the route calculator works. It is also not completely known how changes in the algorithm affect the calculated routes. The estimated risk when the delivery is not in time is acceptable, as Statistics Netherlands saves estimated routes. These "saved" routes can be used as a fall-back scenario. When an update is not delivered in time, the previous release can still be used. As changes in roads will only affect a small part of the routes and changes in route lengths will most of the times be limited, working with an old release will most of the times not lead to large differences.

The conceptual *metadata* related quality issues in the Metadata hyperdimension all scored good as the route calculation tool uses the same definitions as Statistics Netherlands. Also the population of the units used by the route calculator, e.g. all places between which routes and corresponding lengths can be determined, does not lead to problems as all routes that are of interest to Statistics Netherlands can be calculated. The reason for scoring the knowledge about the *data treatment* with reasonable is that Statistics Netherlands does not know exactly how the algorithm used by the route calculator works and how any corrections are made.

5.2. Scores internet prices

Determining the scores for the *supplier* dimension turns out to be difficult in the case of internet prices. The *supplier* of internet prices can in fact be interpreted in two ways: the web host or the company whose prices are displayed at the website (these can be different). Another difficulty is that when prices are gathered by visiting the website (manually or automatically) there will -in general- be no direct contact with the supplier[1]. Although the owner of the website will

[1] In general contact information can be found at websites, it is very questionable however whether this 'contact' can be used by Statistics Netherlands to ask questions.

most likely be informed about the use of the website. 'Most likely' is written as this topic and the way to communicate with the "data supplier" is currently under investigation (Hoekstra et al., 2010). This direct contact is however assumed by the quality framework.

At first glance all types of internet prices score positively at the *privacy and security* dimension. Since all collected information is publicly available this does not seem to be an issue. No private information is exchanged. This is however currently investigated by the legal department of Statistics Netherlands making a distinction between non-substantial and substantial use of a website (Hoekstra et al., 2010). For substantial use, signed permission of the data owner seems required.

When Statistics Netherlands collects price information from a website this can be interpreted as a kind of *delivery*. This is however a completely different type of delivery than in the case of registers. The fact that there is no real delivery implies that the delivery related quality aspects can be scored positively. The only interesting part is whether all information needed by Statistics Netherlands can be delivered.

Regarding the *procedures* it can in the case of internet prices be generally stated that it does not matter for Statistics Netherlands how prices are determined. When the prices mentioned at internet are the prices the customer actually pays, no problems occur.

More attention needs to be paid to the quality indicator "familiarity with planned changes of data source" (see Table 1). This has two reasons. For one, when Statistics Netherlands uses dedicated software to collect prices from a given website, problems can occur when the website changes (Hoekstra et al., 2010). This is however only a minor problem in the case of manual data collection (as is currently the case in producing the CPI). For two, the frequency of visiting the website is directly related to the frequency of changes in the prices at the website. At the website regarding the housing prices the visitor can select a set of houses for which he gets a message when there are any changes at the website. Information about planned changes in the structure of the website is generally not found on websites.

As there is no real contact between the "supplier" and Statistics Netherlands it is also difficult to communicate with the "supplier".

Regarding the last quality indicator belonging to the procedures dimension (Table 1), i.e. the availability of fall-back scenarios, it can roughly be stated that these scenarios are needed in two cases: (1) a program that automatically collects prices does no longer work. This case, for instance, occurs when a web host changes the layout of its webpage's, (2) the website is not accessible. In the first case the problem can theoretically be solved by temporarily collecting the data manually. The word "theoretically" is written in this context as automated data collection can in the future offer many new opportunites for creating, for example, new statistics requiring large amounts of data. When the amount of data becomes to large manual collection will be no longer possible within a reasonable time period. In the second case the "real" shops can be visited given that these do exist and use the same prices (see "comparability" dimension). The latter is, for example, the case for the supermarket. However, if the prices on- and offline are not the same the only possibility is to wait until the website is available again.

Regarding the *clarity* of the metadata it can generally be stated that the meaning of the variables presented at the websites was clear. At the supermarket site product prices, the corresponding quantities and the price per "standard quantity" were provided. For the flight tickets in general a clear price specification was found: ticket price, taxes, reservation costs, and so on were most of the times clearly separated. If this was not the case it was at least stated that the presented prices included all taxes and so on. For the supermarket website and the website with price information of houses problems arose in determining the definition of the population. For the supermarket website it was not completely clear what the whole assortment of the online shop was. The supermarket chain distinguishes several different types of 'real world' shops, that each have a slightly different assortment. The website of the online 'supermarket' did not contain any information on the type of shop and hence the assortment. At the website with house prices it was mentioned that not all houses that are for sale were shown at the investigated website. More specific

only houses sold by real estate agents connected to several unions could be shown at the website, but is was not clear whether all these houses were actually shown.

For the websites it generally holds that it is difficult to determine the time dimension to which the prices refer. The prices are the prices at the moment the website is visited; it is not clear how often or when these prices change. A positive exception is the house prices website, here a user can get a message when the prices of selected houses change. At the site of the supermarket clear information is provided about the time period during which special offers can be bought.

The *comparability* of the definitions used on the websites to the definitions of Statistics Netherlands could often not be determined. This is due to the fact that Statistics Netherlands does not use this information currently in producing statistics. For the websites regarding flight tickets no problems regarding the comparability were expected to occur. Regarding websites it should of course in general be questioned whether the products for which the internet prices are displayed do have exactly the same price in "real world" shops. This question could be answered for the supermarket website, at this website it was stated that the prices are equal for the web and the "normal shops". Note that "unequality of prices" is not necessarily a problem as shopping via the internet becomes more and more popular, so when a statistic would be produced related to prices in webshops there would be no problem. It only is a problem when the prices are used as being representative for prices in "real shops".

Unique keys for the products could be found on all websites. For example, for the houses complete address information, including postal codes, was provided. For the unmanned filling stations, products could be uniquely identified by the type of fuel and the location of the filling station. For the flight tickets a combination of variables can be used as primary key (this combination could include, for example, starting location, destination, time of departure, time of arrival, type of flight (return trip versus one way), type of rate (economy versus business class), and company). For supermarket prices, the products at the website were well described and a picture was provided making it most likely possible to identically identify the products (although a product code was missing).

The way in which the *data are treated* is not really of interest given that customers pay the prices mentioned at the website. The question that should therefore be included here is: how likely is it that the prices on the website do contain errors?

6. Conclusions about the applicability of the quality checklist

In this section we discuss whether the quality dimensions distinguished in the first two hyperdimensions of the quality framework for registers turned out to be useful in determining the quality of offline route information and internet prices. In Table 5 we consider the applicability of the quality dimensions related to the Source hyperdimension. In Table 6 we consider the applicability of the dimensions regarding the Metadata hyperdimension.

A score in these tables can be interpreted as the "extent" to which the specific dimension is useful for a specific data type. The use of the symbols is different from Table 3 and Table 4: relevant (+), partly relevant (o), generally not directly applicable (-).

In general it can be concluded that the majority of the dimensions contained in the Source hyperdimension do also apply to the offline route calculation tool (Table 5). The reason is that also for the route calculator there is a supplier that is known to Statistics Netherlands and that frequently delivers information (updates of the software and the maps used by the software). The only difference is the "privacy and security" dimension as in fact no private data are exchanged between the data supplier and Statistics Netherlands. But this can also be the case for registers not containing private information.

The use of internet prices shows some inherent differences with the use of registers leading to difficulties in applying the Source hyperdimension part of the framework (Table 5). An important difference is that the information is not directly delivered to Statistics Netherlands by the "supplier" (although it is currently investigated in which way the 'supplier' needs to be informed). Statistics Netherlands rather collects the information itself by visiting the website manually or automatically.

This makes the supplier and delivery dimensions less applicable to internet data. The privacy is at first glance also not an issue as all collected information is publicly available at the internet; legally this is somewhat more difficult (Hoekstra et al., 2010). The "procedures" information is also less relevant: when prices at the internet are indeed the prices a customer pays it is not important for Statistics Netherlands how these prices are determined.

As using data requires exact knowledge about the content of these data, the quality indicators in the Metadata hyperdimension are relevant for registers as well as for offline route information and internet prices (Table 6). The only difference refers to the data treatment by the supplier in the case of internet prices. As stated before, when the prices on the website are indeed the prices a customer pays it is not important for Statistics Netherlands how these prices are treated by the owners of the website.

Table 5 Applicability of the quality checklist for the Source hyperdimension

	Offline route information	Internet prices
Supplier	+	-
Relevance	+	+
Privacy and security	o	o
Delivery	+	-
Procedures	+	o

Table 6 Applicability of the quality checklist for the Metadata hyperdimension

	Offline route information	Internet prices
Clarity	+	+
Comparability	+	+
Unique keys	+	+
Data treatment	+	o

6.1. Missing quality indicators

In the previous sections we considered whether all dimensions evaluated within the checklist do apply to offline route information, and internet prices. In this section we use the experiences obtained while filling out the checklist to list some missing quality aspects:

- *Availability of the website* As Statistics Netherlands will collect internet prices by visiting the website either manually or automatically it is important that the website is available when Statistics Netherlands needs the information. The percentage of time during which the website is available could therefore be an important quality indicator for a website.
- *Burden for website* When Statistics Netherlands starts to collect data automatically from a website, this could cause a large burden on the internet traffic of the corresponding website. In this case it is well imaginable that Statistics Netherlands decides to collect the data when the website is not visited a lot. For example, during the night. It is therefore needed to consider how large the burden is for the website, and how (and whether) Statistics Netherlands can assure that no problems are caused to the availability of the website.
- *Errors at website* When using internet prices it is important to know if and how often a website contains errors meaning that customers actually pay prices different to the ones displayed on the website.
- *Possibility for automatically collecting prices* When Statistics Netherlands wants to collect the price information automatically it also needs to be considered whether the lay-out, and technical composition of the website allow this.
- *Representativity of the website* Another interesting question is how representative the website is for the information that needs to be obtained. For example, how much revenue is obtained via the website? Are the prices at the website representative for the prices in "real" shops? Does the "Supplier" apply a central pricing policy? Note that this indicator can well be a special component of the "Comparability" dimension.

7. Conclusions and future research

In this paper we discussed whether the quality checklist for registers can also be applied to offline route information and online prices.

In general it can be concluded that the quality checklist is a useful tool in determining the quality of the offline route information. This can be explained by the relatively large resemblance between registers and offline route information. Both data sources have a clear supplier, and this supplier delivers the data to Statistics Netherlands.

In applying the framework to internet prices several problems arose regarding the "source" dimension. The main reasons were: when collecting prices from websites there is no real supplier (as is assumed in the checklist), there are no real deliveries, and the procedures do not really matter as long as the displayed prices are the prices the customer pays. The part of the framework related to the Metadata turnes out to be useful.

For internet prices we could already point at four missing quality indicators: the availability of the website, the burden for the website from which the data are collected, the number of errors at the website, the possibility for automatically collecting prices and the representativity of the website.

The general conclusion from these results is that the part of the quality checklist applying to the "Source" is most likely only useful for data sources having a clear supplier and clear deliveries. The Metadata part turned out to be useful in general. This is no surprise as working with data always requires good knowledge about the data definitions.

Given these conclusions more research is needed to develop a framework for the Source hyperdimension of internet data including at least the above mentioned "missing quality indicators", and excluding the quality indicators not found to be applicable. To change the Supplier dimension such that it becomes applicable to websites it needs (among others) to be investigated whether and how the owner of a website needs to be informed about the use of a website. And whether and how contact with the supplier is needed and possible.

8. References

Daas, P. J. H., Arends-Tóth, J., Schouten, B., and Kuijvenhoven, L. (2008), Quality framework for the evaluation of administrative data.In: *European Conference on Quality in Official Statistics 2008*, Rome, Italy.

Daas, P. J. H., Ossen, S., Vis-Visschers, R., and Arends-Tóth, J. (2009), *Checklist for the quality evaluation of administrative data sources (Discussion paper)*. Report 09042, Statistics Netherlands, Heerlen.

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

Hoekstra, R., ten Bosch, O., and Harteveld, F. (2010), *Automated data collection from web sources for official statistics: First experiences*. Statistics Netherlands, Leidschenveen.

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

Roos, M., Daas, P., and Puts, M. (2009), *Waarnemingsinnovatie: nieuwe bronnen en mogelijkheden (Discussion paper)*. Report 09027, Statistics Netherlands, Heerlen.