Comparative assessment of three quality frameworks for statistics derived from big data: the cases of Wikipedia page views and Automatic Identification Systems

Fernando Reis[1], Loredana di Consiglio[1], Bogomil Kovachev[1], Albrecht Wirthmann[1], Michail Skaliotis[1]

[1] *Eurostat, Luxembourg, Luxembourg; albrecht.wirthmann@ec.europa.eu*

Abstract

National and international statistical agencies are currently experimenting with the production of statistics derived partly or entirely from big data sources. At the same time there have been initiatives in the official statistics community and elsewhere to extend existing quality frameworks to statistics whose production involves the use of this type of data sources. UNECE's suggested framework for the quality of big data and Eurostat's accreditation of big data sources as input data for official statistics are two examples in this regard. The framework proposed in the report on big data of AAPOR (American Association for Public Opinion Research) is an example coming from outside official statistics. These frameworks have been developed based mostly on theoretical considerations, even if early experiments have provided some input. In this paper, we propose to enrich the experience in the application of these frameworks to particular use cases of statistical products based on big data sources in order to assess their suitability, feasibility and completeness. We apply these three quality frameworks in the context of "experimental" cultural statistics based on Wikipedia page views and to data from Automatic Identification Systems (AIS) for the production of transport statistics.

Keywords: big data, official statistics, quality, Wikipedia page views, AIS.

1. Introduction

Recognising the opportunities and challenges posed by big data, the heads of the national statistical offices in Europe agreed in Schevenningen, in the Netherlands, on a memorandum to address the introduction of big data in official statistics.

Following the Schevenningen memorandum, an ESS task force on big data was set up with the purpose to develop an action plan and a road map for the integration of big data in official statistics in Europe and subsequently to follow those up. The strategy of the roadmap was to start with an initial version of the action plan and at the same time launch some pilots making use of big data sources for official statistics purposes in order to gain a better understanding of the implications of big data for official statistics. Based on the lessons learned from those pilots, the actions would then be reviewed in further revisions of the action plan and roadmap. Pilots have been launched at several national statistical institutes and at Eurostat, individually and jointly at European level. One of the pilots initiated by Eurostat and then run in the context of the big data sandbox initiative promoted by the UNECE, was the use of the data on the use of Wikipedia, now presented in this paper.

2. Big data

Big data is a term which has been used with many different meanings. In this paper, we understand big data as the digital trace left unintentionally by people during their use of computer mediated systems or captured by sensors. This digital trace is often very detailed and of very large size.

The list of possible big data sources is potentially limitless. UNECE (UNECE, 2013) proposes a classification of big data sources based on how they are generated. Firstly, human-sourced information available mostly from social networks where data are loosely structured and often ungoverned (e.g. Facebook and Twitter); Secondly, process-mediated data available from the IT systems of enterprises, where data is usually structured and stored in relational databases (e.g. credit card transactions stored by banks); Thirdly, machine-generated data captured by sensors and other machines used to measure and record events in the physical world (e.g. traffic sensors and web logs). Big data sources potentially relevant for official statistics are those which cover large portions of populations of interest and which can potentially provide answers to questions raised by policy makers and the civil society.

As a new data source, big data offers great advantages and challenges for official statistics. On the one hand, big data sources offer the possibility of higher timeliness, increased efficiency of statistical production systems (which is not to say that big data sources are completely costless), much higher detail and significant development in regional and spatial statistics. Big data sources consist of direct measurements of phenomena and not on indirect reporting by a survey respondent, which may result in improved accuracy.

On the other hand, as single big data sources cannot be expected to answer all statistical needs and will need to be combined with survey data, administrative sources and other big data sources, the complexity of production systems will increase. One important challenge of big data sources is the assessment of the quality of the statistics with it. They very often suffer from selectivity, which may lead to biased results if one does not measure and account for it. However, they stress in particular the existing quality frameworks in statistics by bringing new factors which were not present in traditional sources.

3. Statistical quality frameworks

Currently well-established quality frameworks assumed surveys as the main data source. Hence indicators for quality dimensions were often developed having these in mind. More recently, with the introduction of administrative data in statistical production, it was necessary to adapt methods for quality assessment which led to developing quality frameworks for administrative data. Administrative sources share some characteristics with Big Data. However, they are generated in a more controlled environment (often based on legal requirements with defined characteristics, monopoly in data collection, etc.) while the generation process of big data is not controlled. Therefore, recently several new statistical quality frameworks have been developed, either dedicated to big data or taking big data into account.

3.1 UNECE suggested framework for the quality of big data

The United Nations Economic Commission for Europe (UNECE) Big Data Quality Task Team, set up in the context of the project "The role of Big Data in the Modernisation of Statistical Production", suggested a framework for the quality of big data (UNECE, 2014). It was based on existing quality frameworks, thus inheriting their main characteristics. It addresses quality issues following an input - throughput - output model for statistical production and adopts a hierarchical structure, as already suggested in the quality framework for administrative data, where quality dimensions are nested in three hyperdimensions: source, metadata and data. The data hyperdimension relates to the data themselves, whereas the source and metadata hyperdimensions relate to the conditions that govern the data supply and to the availability and kind of information available on the concepts and contents, respectively. The three hyperdimensions and corresponding quality dimensions are considered in each phase of the input - throughput - output model (see table 1).

Table 1: *Quality dimensions in the UNECE suggested framework for the quality of big data*

	Input	Throughput	Output
Source	Institutional/business environment	System independence Steady States Quality Gates	Institutional/business environment
	Privacy and Security		Privacy and Security
Metadata	Complexity		Complexity
	Completeness		
	Usability		
	Linkability		
	Coherence - consistency		
	Validity		
	Time-related factors		
			Accessibility and Clarity
			Relevance
Data	Accuracy and selectivity		Accuracy and selectivity
	Linkability		Linkability
	Coherence - consistency		Coherence - consistency
	Validity		Validity
			Time-related factors

In this quality framework, specific aspects related to big data are taken into account by considering a new quality dimension (in comparison to proposed frameworks for the quality of administrative data) for the complexity of the input data, which the need for new skills and new IT infrastructures.

It is possible to use the framework to evaluate input quality even before data is actually available. In such cases only the hyperdimensions source and metadata are concerned. At this stage it is usually the potential of the source that is assessed both relating to existing indicators and new outputs. When data becomes available metadata quality can be re-assessed, analysing in detail the coherence between the description and the observed data values.

Regarding throughput, which concerns the processing of the data, the quality the framework is not concrete in terms of indicators and describes some general principles instead, system independence, existence of steady states and existence of quality gates. System independence means that data processing should follow theoretical principles and not be dependent on the system that implements them. Steady states consist of intermediate datasets which are maintained in the data processing pipeline for control purposes. Quality gates are points in the data processing pipeline where data quality is assessed.

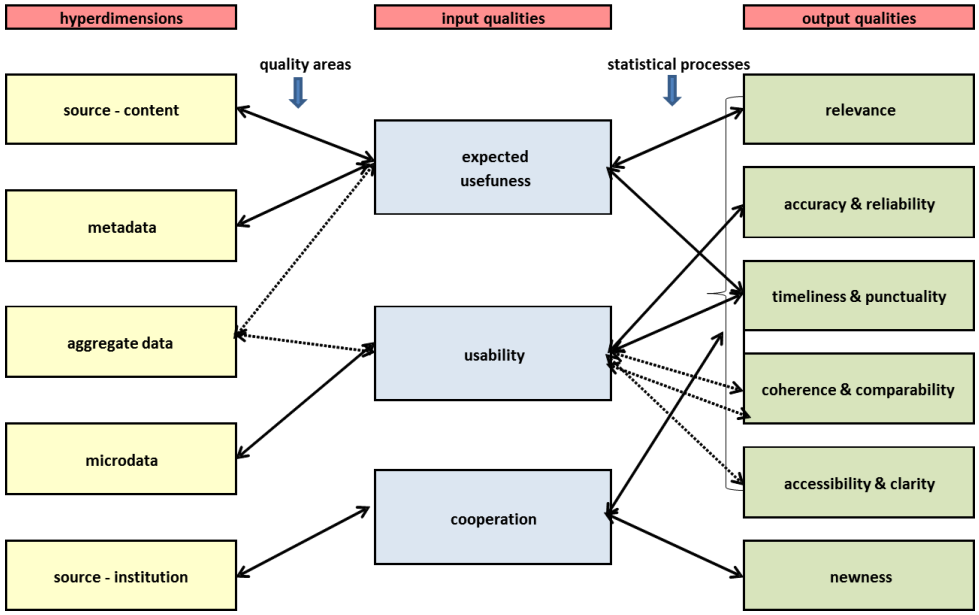
3.2 Eurostat suggested accreditation of big data sources as input data for official statistics

In the context of Eurostat study "Analysis of methodologies for using the Internet for the collection of information society and other statistics", an accreditation procedure for statistical data from non-official sources was developed (Eurostat, 2013). This accreditation procedure included a quality framework which was based on existing frameworks directed to the use of administrative sources for statistical purposes, extending it to any type of secondary source.

It uses the concept of hyperdimension, linking it explicitly to objects subject to quality assessment while quality dimensions addressing characteristics of those objects. It starts from hyperdimensions used in other quality frameworks, and splits the source between the institution that holds the data and the content, and the data between aggregate data and micro-data. Furthermore, it avoids confusion by defining input qualities with terms which do not overlap with those defined for the output. Finally, it takes into account the particular uses of the quality framework, by grouping the traditional output quality dimensions into one for those cases where the impact of the use of a new data source is assessed in existing statistical

outputs and distinguishing the case of completely new statistical outputs which would have a new quality dimension named newness.

Table 2: *Hyperdimensions and qualities of inputs and outputs in the Eurostat suggested accreditation of big data sources as input data for official statistics*



The accreditation procedure is a step-wise approach with five stages evaluating incrementally the quality of a source. At each stage a decision on whether to proceed has to be taken.

A first evaluation is done with very basic information on the content of the source and the metadata. The first stage aims to assess the potential usefulness looking at coverage, units and variables, timeliness and frequency. At this stage, data is not strictly necessary.

The second stage relates to exploring the possibilities for obtaining the data. This includes discussions on the necessary aggregation level and format for exchange.

It is with the third stage that the NSI start analysing the quality of the data, at this stage for explorative purpose. The stage is subdivided in four sub-phases: first data quality assessment and processing (duplicate records, validity and consistency of the recorded values, treatment

of the outliers), then after a decision point, the data is cleaned, imputed and weighted and then the final data to be used for estimation is produced. Indicators at this stage can be compared with existing data. As a final sub-stage, the evaluation of new products is carried out together with the assessment of the available technical tools to treat the big data source.

Stage 4 entails the NSI's decision on the acquisition of the source: including the analysis of new outputs, implications, trade off and impact on timeliness. A cost-benefit assessment together with a risk evaluation on the consequences of factors not in control of the NSI with the mitigation strategies to put in place is carried out before to proceed.

In the process of evaluation of a new source, the data quality assessment is joint with the financial, but also with the legal aspects of acquiring an external source and the risk management issues. The final stage is the actual agreement and cooperation between the NSI and the data provider.

3.3 AAPOR total error framework for big data

The AAPOR total error framework for big data (Kreuter, 2015) addresses specifically the accuracy of the output. It extends the total survey error (TSE) to big data mainly to include new possible sources of errors and uncertainty that are typical of the generating and manipulation processes of big data sources. TSE is already based on the description of the sources of errors, subdividing the causes of bias and variance components.

Similarly, the Big Data Total Error (BDTE) assumes the presence of row errors (omission, duplication or erroneous inclusion of an "element"), column errors (e.g. wrong definition of variables) and cell errors (including content error, specification errors and missing data).

Omission of row element is a common error in big data, causing bias if no adjustment is made for it. A typical case of erroneous inclusions is an object associated with a computer (e.g. in the case of queries in a search engine), instead of an individual of the population. Column error can occur in the case of big data sources when these are subject to many transformations (e.g. machine learning) before being usable for statistical production.

The analysis of the process generating the errors in a survey (Biemer, 2010) and in administrative data (Walgreen and Walgreen, 2007) is very well developed and the AAPOR report extends the analysis to big data. In spite of the difficulty to generalize to sources that may in fact greatly differ one from another, it illustrates three generic steps where errors may originate: generation, ETL (extraction, transform and load) and analysis.

In the first step errors relate to the data generation process and include high noise to signal ratio, lost signal, incompleteness, non-random selective source, and missing metadata.

In the second step errors relate to the processing of data including errors in specification, matching, coding, editing, data mugging and data integration..

In the third step error relate to the analysis of big data including noise accumulation, spurious correlation and incidental endogeneity.

4. Application to two cases

4.1 Wikipedia as a data source

Wikipedia was founded in 2001 by Jimmy Wales and Larry Sanger with the objective of creating a free online encyclopedia that anyone can edit. In the last 15 years it has grown to 38 million articles in 246 languages. It is widely used with 21 million page views per hour. According to European official statistics, in 2013, 44% of individuals of 16 to 74 years old living in the EU consulted wikis to obtain knowledge (e.g. Wikipedia). This was 69% for individuals between 16 and 24 years old.

While using Wikipedia, people leave digital traces of their activities, in particular as a result of accesses to and editions of Wikipedia articles and their corresponding discussion pages. These digital traces exist as data in the web logs of the servers which host Wikipedia and in the content of the articles and their corresponding discussion pages.Detailed data on the number of page views per article is made publicly available by the Wikimedia Foundation - the organization which supports and hosts Wikipedia.

Apart from the raw number of views there is additional information available such as the language version of the article, its textual and graphic content, its human assigned classification, its edit history (which allows the identification of controversies, e.g. as evidenced by the so called edit wars). Of particular importance for UNESCO World Heritage Sites are the so called information boxes (infoboxes) that provide structured detailed information.

For our analysis we have used two of the components above. The first one is the number of page views which is made available as dump files in several different formats. A page view in these datasets is defined to be the number of time the page has been accessed not counting access via mobile devices and access identified as done by non-humans (i.e. bots). It is also corrected for outages (i.e. when the Wikipedia servers are not available). The second component is the content of the articles. It is readily available and we have obtained it via the Wikipedia API.

Quality assessment with the three frameworks

The application of the quality framework proposed by the UNECE task team to the input quality allowed identifying the strength of the "institutional environment" of the Wikistats, supported by a non-profit organisation (Wikimedia Foundation) with high standards of transparency, and high level of sustainability given the Wikipedia is open content. Other strengths identified were in the "time-factors", namely its timeliness and high level of temporal detail (hourly data) and the "completeness" of the metadata. It also helped identifying some weaknesses, such as the "complexity" of the input datasets, which require some specialisation in an office which would like to use this data source and also on the "completeness" of the metadata which although very detailed was difficult to navigate and at some points not clear.

One of the lessons we may take for the framework itself from this case study are the unclear implication of the sustainability of these big data sources for input quality. Although Wikistats do not suffer particularly from lack of sustainability, most big data sources do. However, does

this mean that because a data source may not be available in the future then we should not use it, even if meanwhile it may provide very valuable statistics? The relevance of such question depends on the use of the quality framework: decision to acquire a new data source, comparing alternative data sources or reporting. In the latter two cases it is straightforward how to consider sustainability, but in the former the answer depends on the impact on the output quality. If the relevance of the statistics is also limited in time (e.g. some information society statistics) then the lack of sustainability of the input plays a lower role.

The quality dimension "privacy and security" was difficult to apply to the Wikistats case, as it did not involve personal data. In fact, all the factors seem to be linked to the "institutional environment", mostly of the statistical office, and probably they should be together with that quality dimension. The quality dimension "complexity" should distinguish the nature of the data itself from the data formats available. In the case of Wikistats, even if the file format was complex, the institutional setup is such that file formats can possibly be negotiated with the Wikimedia Foundation and complexity reduced.

In terms of throughput quality, the three general principles proposed, system independence, existence of steady states and existence of quality gates, were very useful. Our processing of Wikistats was based on open source tools and open algorithms (available via GitHub). We did have steady states, as raw files made available by the Wikimedia Foundation were pre-processed and then available for a multitude of uses and analysis. However, we wonder how feasible this will be for other big data sources where datasets are even more massive than the ones we used or rapidly changing, as it's typical of some big data. Other forms of auditability may need to be used, such as historical samples and log files. We did not have quality gates defined a priori, as this was a pilot, but they seem applicable to our use case.

The output quality factors and indictors were more difficult to apply to assess Wikistats. Many did not add much to the factors assessed for the input and would only be relevant if using the framework for reporting. This is in fact pointed out by the UNECE big data quality task team in their report. Quality dimensions which added less to the assessment of the WHS statistics produced with Wikistats were "institutional/business environment" and "privacy and security"

where it mostly referred to the input. It could include some new elements related to the environment of the statistical office producing the statistics which would be very important for quality reporting. "Complexity" also referred to the input and its implications for the output. Beyond reporting on the treatment of the input, the complexity of the output, e.g. the existence of nested hierarchies, will be reflected in other dimensions, positively on relevance, if that complexity brings analytical richness to the data, and negatively on clarity.

Overall the only quality factor which seems to be missing, and existing in other quality frameworks, and both in the input and output, is punctuality, the difference between the planned and effective time of data release. This could be introduced among the time factors.

The application of the total error framework for big data to the analysis of the output accuracy of the WHS statistics based on Wikistats was very useful to identify sources of error and to make the link between input, throughput and output. In the data generation step, it allowed identifying the crowdsourced nature of the Wikipedia articles as the main source of noise and the constant update of the Wikipedia as the main source of lost signal (e.g. deleted articles disappear from the Wikipedia, including their history).

In the ETL step, specification error seemed more evident when a study is designed and less when data products are built on exhaust data. However, upon reflection it turned out to be applicable to our use case. Specification error is possible when selecting the articles relevant for the topic of interest (in our case the articles related to world heritage sites), or selecting the categories (i.e. article categorisation made by wikipedians) of articles to study, or even when constructing relative indicators of page views where several different bases can be used (e.g. page views of the "Main Page" article or the total page views of the language version of Wikipedia). Other sources of error were also identified.

In the analyse step, the sources of error mentioned in the AAPOR report were not applicable to our use case because they refer to statistical modelling which we did not apply. However, the use of model based inference will most probably be more prominent with the increasing use of big data.

The framework developed by Eurostat for the accreditation for statistical purposes from non-official sources didn't bring additional quality dimensions that we did not find in the previous ones. However, it provided a clearer overall framework, distinguishing the output quality of the input data from the input quality. The quality of the input data would be accessed assessed using the output quality dimensions, while the input quality would be composed of other dimensions relevant for the assessment of a data source as an input, namely the expected usefulness, the usability, and the institutional environment (there named "cooperation"). It also links that framework to a specific process of decision making concerning the acceptance of a new data source. The input quality dimensions listed in the UNECE proposal can all be mapped to these three input qualities. What the accreditation framework is missing is explicit quality factors for the throughput.

4.2 Automatic Identification Systems (AIS) as a data source

AIS is primarily a safety instrument required by the International Maritime Organization's (IMO) International Convention for the Safety of Life at Sea (SOLAS) that became fully operational in 2008.

It provides a means for ships to electronically send data (about their position, destination, speed, etc.) to Vessel Traffic Services (VTS) stations as well as to other nearby ships. It is used for managing vessel traffic and avoiding vessel collisions.

AIS uses a positioning system, such as the Global Positioning System (GPS), in combination with other electronic navigation censors and standardised Very High Frequency (VHF) transceiver to automatically exchange navigation information electronically.

AIS messages are transmitted by ships using VHF signals. Vessel identifiers such as the vessel name and VHF call sign are programmed in during initial equipment installation. These are included in the signals transmitted by vessels along with location information originating from the ship's global navigation satellite system receiver. By transmitting a signal, vessels can be tracked by AIS base stations located along coastlines.

When a vessel's position is out of the range of the terrestrial networks, signals are received via satellites. When the combined coverage of the ground stations and satellite receivers is sufficient, AIS is capable of globally monitoring a ship's movement in real time.

AIS is obligatory for vessels over 300 gross tonnage (GT) on international voyages, and for all passenger ships and vessels over 500 GT on domestic voyages. However, a very large number of vessels are fitted with AIS and the number is growing as smaller and cheaper devices are fitted even in small vessels on a voluntary basis.

AIS transceivers send data every 2-10 seconds depending on the vessel's speed-or every 3 minutes if at anchor. Basic information are latitude, longitude, Speed Over Ground (SOG), Course Over Ground (COG); (and every six minutes: identification (IMO and MMSI number, ship name and call sign), static (size, type of vessel, type of cargo, etc.) and voyage related information (e.g. Estimated Time of Arrival and destination).

These data is collected by maritime traffic control and then sent to the European Maritime Safety Agency (EMSA) . EMSA's SafeSeaNet (SSN) integrates AIS data with other sources. In this paper we are focusing on the data obtained from EMSA. Getting data directly from maritime traffic control may sometimes also be possible. However the quality evaluation of the source would possibly be different in such circumstances.

Quality assessment with the three frameworks

In the case of AIS, we still do not have access to the data. However, some preliminary assessment on the potential use of the data source is possible as both the Eurostat proposal for an accreditation and the UNECE suggested quality framework have a stepwise model for the evaluation in phases and a input - throughput - output model.

On the other hand, a reflection on the kind of errors affecting the BDTE is still possible even without having the possibility to measure the impact of the single components of the TE.

Both the accreditation and the quality framework introduce the need for evaluating the reliability of the data provider to assess and evaluate the risk of discontinuity in the data source, with possible impact in the data production, and as cause of time-inconsistency. The UNECE framework includes at the input stage the identification of potential issues in legacy legality and confidentiality, whereas the accreditation only assesses it after having evaluated the potential usefulness for official statistics.

AIS are used on the basis of a safety requirement by the International Maritime Organization's (IMO) International Convention for the Safety of Life at Sea (SOLAS), from 2008 and a procedure to request data is made available on EMSA website.

The source might undergo into technical changes (e.g. the AIS signals received by satellite, a system already implemented and in view of improving the geographical coverage of transmission). However, changes and transmission methods are documented.

The description of the data source on EMSA website permits identifying the units and variables as well as the frequency of updates. Moreover linkage variables, namely the ships identifiers, are present and actually used by EMSA to enrich the AIS data with information on the type of vessel.

An assessment of usability, for this source, in terms of need for acquisition of new resources is related to the type of desired output and to the level of aggregation of the data to be stored. EMSA itself recognizes the potential of the source to produce statistics on the numbers of ships; the different ship types; numbers of journeys; numbers of accidents and types.

In what concerns the quality related to data, with the information at hand, only a preliminary assessment of the possible causes of errors to decompose the BDTE is possible without measuring. In particular it is possible to envisage the possibility of under-coverage (omission of vessels not under regulation), missing values (data not transmitted) and measurement errors (erroneous position). Finally, if data are reduced to facilitate the analysis, wrong models may impact the conclusion drawn.

5. Conclusions and lessons learned

All three frameworks used in this paper were useful and they mostly complement each other. The UNECE one provides a comprehensive and detailed list of factors and indicators which allowed assessing the data sources and the statistics which could be produced. We could fit all issues we had previously identified in an ad-hoc manner while working on the project and we found additional ones we hadn't thought of. On the other hand, it was less clear on the application of the framework with some indicators which seemed more applicable to a decision to acquire, or not, the data source and other which made more sense only in the context of a quality reporting.

The AAPOR total error framework for big data is very detailed and provided a clear link between input, throughput and output. This is important for identifying sources of error, but also to assess the implications of input quality for the output quality, which is very important if a quality framework is to be used to decide on the acquisition of a new data source or to choose between data sources. However, it is restricted to the accuracy quality dimension. Similar dedicated frameworks could be imagined, for example, for timeliness (very relevant for big data sources).

The accreditation framework developed by Eurostat is very clear in the application of a quality framework to the acquisition of new data sources and on the impact of input quality in the quality of the output. It also distinguishes the output quality of a data source which we may use as input and its quality as an input to a particular statistical production process. In our view, this is a very important distinction. For example, the output quality of a statistic in terms of relevance may be very good because it provides very high detail which is needed for its users, but it may have a lower quality as an input to our process of interest because it is too complex for the statistic we want to produce. However, this framework is very specialised to the assessment of new data sources, lacks throughput quality dimensions and does not provide the detail found in the previous two frameworks.

Judging from the case studies here presented, elements of each of these frameworks could be combined to produce a better quality framework. This eventual improved framework, while combining the strengths of each of them, should clarify its use for assessment of the acquisition of a new data source, comparing alternative data sources, assessing the impact of changes in input data sources and throughput processes and in quality reporting. It should also make clear the links between input and throughput quality and output quality.

The first and third frameworks are based on frameworks which had in mind the use of administrative sources for statistical purposes. Most of the quality dimensions identified having administrative sources in mind were indeed applicable to our use cases, following the indicators provided by the adapted frameworks analysed in this paper. A legitimate question is if the existing frameworks, which take administrative sources into account, are enough to address big data. Our case studies have shown that new quality dimensions, such as complexity, are indeed important and that the new quality indicators proposed are needed.

6. References

Biemer, P. P. (2010). *Total survey error: Design, implementation, and evaluation*. Public Opinion Quarterly, 74(5), 817-848.

Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., ... & Usher, A. (2015). *AAPOR Report on Big Data* (No. 4eb9b798fd5b42a8b53a9249c7661dd8). Mathematica Policy Research.

Eurostat (2013) *Accreditation of big data sources as input data for official statistics*

UNECE (2013) *Classification of Types of Big Data*

UNECE (2014) *A Suggested Framework for the Quality of Big Data*

Wallgren, A., & Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes* (Vol. 553). John Wiley & Sons.