Quality Challenges in Modernising Official Business Statistics

Ger Snijkers[*] (Statistics Netherlands), Gustav Haraldsen* (Statistics Norway),
Martin Luppes* (Statistics Netherlands), Piet Daas* (Statistics Netherlands), Johan Erikson*
(Statistics Sweden), and Li-Chun Zhang* (Statistics Norway and University of Southampton).

1. Challenges

The environment for national statistics has changed. A general trend is that statistics are being commercialized and hence that the dominating position of institutes producing official statistics is being challenged. Another trend, more specific to business statistics, is that National Accounts, which is the backbone of National Statistical Institutes, is challenged by the globalization of the economy. Businesses no longer stick to national borders. This trend calls for different (ways of producing) statistics and causes a coverage problem which questions the validity of National Accounts. In this paper we will discuss to what extent improved, but still traditional, data collection methods can meet these challenges, and what new data sources can offer.

Both these trends are linked to the information society, which is here already for some time, but it is deepening, and its effects on society are becoming more and more manifest. It is with the help of modern information and communication technologies that businesses can operate internationally. In addition, more and more businesses are becoming aware of the value of the data they have, data that are collected e.g. as a side product of their processes. The commercialization of statistical products is one example of how data generated by internet activities, by road sensors, or collected from digitalized business systems can be reused and sold.

Likewise, modern information and communication technology also affects the traditional survey data collection methods. Business survey methods need to be adapted to these developments. A keyword here is tailoring survey communication and data collection instruments to the respondents we are addressing.

Computerization has also affected the number and accessibility of registers which can be utilized for statistical purposes. Many National Statistical Institutes (NSIs) have adopted a data collection strategy where data already available should be investigated before the institutes are allowed to launch new business surveys (e.g. Statistics Norway 2007; Statistics Netherlands 2011). The result for statistical institutes is a multi-source/mixed-mode strategy (Snijkers 2009) for getting the required data to produce statistics. The combination and integration of data

---

sources and the mixing of modes in surveys are two important developments we have seen in the last decades. This trend has structural effects on the way work is organized and on the IT-infrastructure in NSIs; related methods and processes are still being developed. In the light of the deepening of the information society this strategy however requires an update: a further specification is needed in order to serve as a blueprint for future data collection designs.

The question we pose in this paper is to what extent access to a wider range of data sources and improvements in data collection methods address present and future needs for high quality business statistics. When discussing this, we think it is important to do so within a quality framework.

2. Qualities

When talking about the quality of data and statistics, there is not *one* quality criteria but rather different quality dimensions. An overall model combining quality dimensions is presented by Haraldsen (Snijkers, Haraldsen, Jones and Willimack 2013: p. 95), called the Quality Diamond, as shown in figure 1. This model is based on Groves' discussion on survey errors and survey costs (Groves 1989), the process approach to Total Survey Errors as discussed by Groves et al (2004: p.48), the Survey Research Triangle presented by Weisberg (2005: p.28), and quality dimensions presented by Eurostat (2011). The main distinction in this model is between a professional quality perspective and the perspective of users of statistics. The professional perspective is depicted as a balance between survey errors, survey constraints, and survey effects. This perspective needs to be balanced with a user evaluation of the data, which is how well the data collected and statistics produced fulfill information needs: how well the statistics are "fit for use". The user perspective can be described by referring to one or more of Eurostat's quality dimensions. New challenges, new data sources, and new data collection methods will certainly affect the quality profile as seen from a professional perspective, how users evaluate official statistics, and the balance between the two. Challenges and developments specific to business statistics are addressed in the next sections.

Figure 1: The Quality Diamond
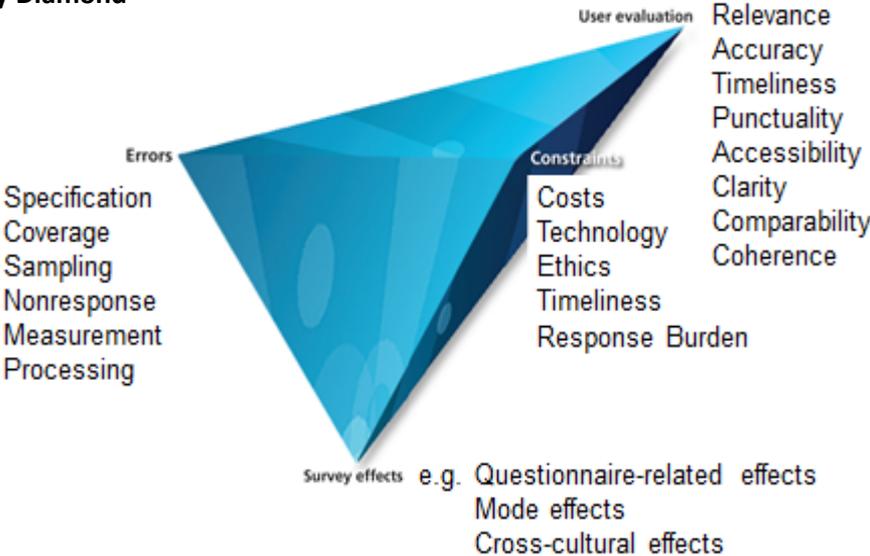
3. Globalization

A major challenge that official business statistics are facing is globalisation. Not only the *broadening* of globalisation, i.e. the increasing number of emerging markets, but also the *deepening* of globalisation in terms of splitting up production processes and its increasing dependency of the international trade in services poses new challenges upon both policy makers and statisticians. As was put in the Eurostat report Global Value Chains and Economic Globalization (Sturgeon, 2013): "We have a strong sense of profound changes in the world economy, […] but cannot fully describe the new patterns and structures […], not least because the official statistics at our easy disposal were created for other purposes and in simpler times." In terms of a user evaluation in our quality framework, this report actually states that official statistics are no longer fully "fit for use"!

Statisticians then face the challenge to measure the growing international interdependency of enterprises in terms of trade, investments, ownership & control and its effects on the domestic economy in terms of jobs and productivity (Van der Veen 2007; Luppes & De Winden 2007; Dicken 2007). This calls for an adapted system of related relevant and accurate statistics, both at the European level (European Statistical System: ESS) and the country level. Already a number of initiatives have been established in this direction like the work within FRIBS (Eurostat's Framework Regulation Integrating Business Statistics) and the Trade in Value Added approaches (Sturgeon 2013; UNECE 2011; OECD 2013).

There are several ways to fill the gaps in the statistical information on economic globalisation. The first step is to combine existing information at the level of the NSI, including the combination of business and social data, using micro-data linking approaches which makes it possible to describe basic parameters of the domestic structures and analyze some of the determinants, its effects and developments (Fortanier et al. 2012). Also the use of secondary sources, including private ones (administrative data, commercial databases, privately collected survey data, etc.) may be increased in order to add information. However, some basic information reflecting the changed production and control structures at the level of the individual enterprise is not available by other means than primary data collection.

Globalisation, due to the process of splitting up enterprises across borders, has far-reaching consequences for primary data collection. In addition to conceptual issues with regard to required information, the conceptual issues outlined in the UNECE report (2011), and also in other work of Eurostat and the OECD (Sturgeon 2013; OECD 2013), reflect the key notion that the statistical unit no longer can be regarded as a black box of which the (aggregated) inputs and outputs form the basis for official statistics. Insight is needed on:
1) the origins and destinations of these inputs and outputs (preferably at micro level);
2) the composition and sourcing of so-called 'business functions' which play an important role in the domestic and foreign dynamic of the labour market;

3) the dynamic structure of enterprise groups which play a key role in foreign direct investments (FDIs), innovation, control and organisational structure of the production chain (governance).

Consequently, depending on the type of governance within an enterprise group or within a given production chain not all relevant information is available at one place. This requires more tailored approaches of communication and combing all data (primary and secondary) available, in order to detect eligibility of (sub)units, and the location of the required information. Also the increasing differences between the conceptual frameworks of statistics and the administrative reality of global entrepreneurship require more and more tailored approaches in data collection. This requires more tailoring of questionnaires and more post-collection processing, since the conceptual differences between the available data and the desired statistical outputs are widening. This can mean splitting up questionnaires reflecting the structure and definitions of the businesses (and not the statistics), and then restructuring the data from business definitions to statistical concepts in the statistical office (perhaps also using secondary sources as additional information in this modelling). Also, profiling units for large businesses within the statistical offices, the implementation of the Eurogroup Register, and the development of data sharing initiatives within the ESS (FRIBS ref) are reflecting this tailored approaches.

If surveys are considered to be a final step in a multi-source/mixed-mode data collection strategy, the final consequence of globalization is that these surveys need to be designed and conducted according to professional standards in order to get "fit-for-use" data (Snijkers et al. 2013). Tailoring the survey design for eligible respondents within the enterprise under study is key in our view. In the next section tailoring of business surveys will be discussed in more detail.

4. Computerizing business surveys

As we have seen, globalization affects primary data collection designs. Globalisation requires rethinking of the concepts that are needed, and consequently the data that need to be collected, especially when using surveys. Furthermore, the arguments made in the previous paragraph call for a different perspective on sampling and frame issues: not only statistical units will change, but also the reporting units, and they need to be opened up. To effectively conduct surveys, the modernization of data collection methods first of all has to be focused on tailoring the survey: tailoring the data collection modes, the questionnaire, and the survey communication (how we address and motivate businesses to participate) to reduce response burden and to get complete, accurate and timely data.

Because much of the information held in businesses is digitalized, and hence much of the work taking place is data processing, and because the internet penetration is high, business surveys are obvious candidates for computerization. Research has shown that the potential for computerized self-administered business survey is far from realized (Gravem, Löfgren et al. 2011). Computerization of the data collection aids to tailor the survey.

One data collection mode that is used or investigated in several countries is to automatically extract the relevant data from business records. This kind of data collection is particularly relevant to businesses that else will have a heavy response burden. The efficiency of automatic data capture, however, heavily depends on a stable and standardized metadata system for business data, a stable structure of business accounts, and stable data requests. Once automatic data capture is established it takes a lot of work to change it. One should also remember that today the most common complaints from business respondents is that the information asked for in business surveys does not match directly with what is available in their records. With automatic data capture this will change; statistical offices will have to collect what is in the business records and transform it into statistical concepts in the post-collection processing.

When questionnaires are used, the flow of question the wording and presentation of individual questions will affect the response quality. The order of sections which is right in the head of the questionnaire designer might not coincide with the structure of the business administration. Because of this the respondent should normally be able to complete the different sections in the order he or she wants. In computerized questionnaires list of topic can be tailored according to relevance decided either by predefined characteristics or filter questions. Next computerized branching can be used to tailor the flow of questions inside each section of the questionnaire.

Individual questions can be tailored in different ways. Two-dimensional matrices, i.e. combinations of units and questions, are common in business surveys. In paper questionnaires these matrices often were presented in an A3 format which had severe quality weaknesses on paper and which cannot be copied on a pc screen. A number of alternative and more dynamic solutions are developed in computerized questionnaires (see Snijkers et al. 2013: p. 353). Features that are particularly useful in business surveys are preloads of previous answers given and running tallies which underpin calculations. Preloading previous answer may, however, result in satisfycing, and thus in measurement errors. A way to deal with this is combining previous answer with error checks.

The most obvious way to improve the response quality in computerized questionnaires is to outsource editing rules so that respondents who break rules have to correct them before the completed questionnaire can be submitted. Too many error checks may, however, add to the response burden and consequently lead to higher monetary and cognitive costs, which also are parts of the quality diamond.

Finally we want to point out that business web questionnaires may be linked to other useful web tools like maps (in transport surveys), classification lists, calendars (identifying reference periods or reconstructing previous events or activities), visualizations and video clips (used for illustrations or instructions).

Irrespectively of the data collection mode used, the appropriate units need to be contacted and informed about the data collection. The survey communication and response process is affected by the business size (Snijkers et al. 2013). In larger businesses the first contact normally is at a

management level, deciding on the priority of the data request, and who should complete the questionnaire. In smaller companies these roles might not be separated. Furthermore, the administrative system in larger business will normally be more complicated than those in smaller businesses: the information asked for will often reside in different administrative databases in different parts of the organization. The information required may also be available from third parties, e.g. accounting bureaus, via which access to many business accounting databases can be acquired. Ideally, business data requests should reflect where different sources of information are located, and these units should be contacted. Finally business surveys often are panels, which means that it is important to accommodate for an ongoing relationship. When business surveys are computerized, one important part of the design is a computerized management system, e.g. a web portal, tailored for case management tasks.

In the light of a multi-source/mixed mode data collection strategy, in which surveys are a final way to collect required data, it should be clear that surveys should be well designed, balancing the survey errors, survey effects, and constraints. From this overview it should be clear that the developments in information and communication technologies offer a lot of options for business surveys to improve the timeliness, accuracy of the data, and reduce response burden. On the other side the survey process will become more complex, and the development costs for collecting designed data will be higher.

5. Quality challenges using secondary data

Rather than improving questionnaires used in business surveys, a different, complementary approach to the challenges posed by new statistical needs, is to expand and improve the utilization of secondary data. As we have already pointed out, more and more NSIs are using secondary sources, like registers, for statistical purposes. With the deepening of the information society, a set of new data sources termed Big Data also becomes available (see e.g. Daas and Puts 2014; Virtual Sprint paper 2014; Glasson et al. 2013; Groves 2013).

The increasing use of secondary sources, results in an increasing awareness of the associated quality challenges. A number of quality frameworks for use of administrative data have been proposed. Daas and van Delden (2013) discern three general approaches focusing on the quality of, respectively, (i) the input data (Daas et al., 2012), (ii) the output statistics (Frost, 2011; Laitila et al., 2011; Burger et al., 2013), and (iii) the metadata (Daas and Ossen, 2011). Zhang (2012) generalizes Groves' total survey error framework (Groves et al 2004) to make it applicable also for register data collected and integrated from multiple sources. With reference to the Quality Diamond in figure 1, one could say that NSIs no longer start from a producer perspective, but rather look at the data from a user's perspective.

There are both similarities and differences between the quality issues raised by registers and Big Data. Even if Big Data per definition is a large amount of data, Big Data sources predo-minantly report events, are often selective and do not cover the entire target population. The same problem is known from registers. The measures that are available in both sources may

also suffer from definitional bias and cause relevance errors if not adjusted. One difference, however, is that it normally is easier to identify the source of bias and measurement problems in registers compared with Big Data. As an example, the selection procedure behind Big Data is often unknown. In this respect the selectivity question is similar to non-probability sampling as is the case e.g. in access web panels (Bethlehem & Biffigandi, 2012).

While surveys are designed to collect data for statistical purposes, registers and Big Data are produced for other purposes (Groves, 2013). As a consequence, extensive uses of secondary data sources make NSIs more dependent on risk factors, often phrased as *preconditions,* as the production of these data are not in their direct control,. UNECE (2007) present five preconditions for administrative data concerning, respective, 1) the legal foundation, 2) public understanding and approval, 3) the availability of a unique identification system, 4) comprehensive and reliable public administrations, 5) cooperation among the administrative authorities (see also Daas and van Delden, 2013). Much resources are used in an effort minimize these risk factors in registers used for statistical purposes, while the preconditions of Big Data are yet to be clarified in some scenarios. When talking about Big Data we distinguish between human sourced information like messages published in social media, process mediated data like business transactions and machine generated data like data collected from traffic sensors. The two latter are the most relevant for business statistics, while the risk factors listed above is most relevant for human sourced information.

In an efficient statistical production process, it is also important to have well-founded, stable and standardized data deliveries. This stability is difficult to establish in a system based on Big Data, one reason being that relevant data science and technical capacity of data processing is lacking in many NSIs.

Because of the large amount of data available in Big Data one might be tempted to process a subset of all the data obtainable. Sampling of Big Data sources, however, may not be straightforward because of the lack of frame, the virtually infinite extension of the net-work structure, or the highly dynamic nature of the 'population'. Moreover, the scheme of selection among the accessible data, the stopping rule of collecting observations that is necessary, and the stochastic nature of the subsequent optimization procedure typical of many statistical methods can all pose new and challenging questions for statisticians.

The large amount of data also comes hand-in-hand with a high proportion of irrelevant and defect (or inadmissible) values. This calls for automatic statistical methods of data editing and imputation in order to deal with the problems that one has previously not encountered when handling sample survey, census or administrative data.

Nearly all Big Data sources are event-based. This is also often true with administrative registers. In the aforementioned extension of Grove's total survey model described by Zhang, he describes the quality challenges affecting the conversion of events to units and identifying and linking the units to the target population. These issues apply to all kinds of data

integration. Because background information often is missing in Big Data (Buelens et al., 2014), it may be particularly troublesome to relate events to units, and consequently to link the units to the target population. A completely different way of looking at this, however, is to question what the most relevant units of business statistics is. In the previous section about globalization we raised the issue of value chains within and across traditional business units. So may be one path to follow is to link events which describe steps in a refining process to units rather than identifying refining processes and the results of these within units.

Another point made in the section about globalization was the need to link social data to business data. It is often not difficult to find Big Data that are apparently correlated with the statistical measure of interest. But new methods of estimation and analysis are usually required to make efficient use of them in order to reach an acceptable level of validity, accuracy and transparency that is customary to official statistics.

6. Modernizing official business statistics

We started this paper by raising the question to what extent modernization of traditional data collection methods and access to new sources of information address future quality challenges of business statistics. In many papers these developments are discussed on its own merits; we have discussed these developments and challenges from a quality framework, and relate those to an integrated data collection approach: the multi-source/mixed-mode data collection strategy. From the discussion a number of basic challenges for modernizing official business statistics can be isolated.

We think it is important to state that the modernization of data collection methods and utilization of new information sources is not as much driven as should be by the need for a new kind of business statistics. The use of secondary data sources and the computerization of data collection and data capture methods should in our view be linked more clearly to statistical needs generated by the globalization of the modern economy. In fact, much of the modernization of data collection methods are not initiated by quality considerations, but rather by budget cuts. We have taken a step towards linking statistical challenges to the modernization of data collection methods by using quality considerations as a common reference point.

The challenges of globalization still need to be specified in more detail before the link with a multi-source/mixed-mode data collection strategy can be made stronger. Only then the potentials of new sources, new data collection and data capture methods can fully serve the modernization of official business statistics. Developing new statistical indicators and an integrated system of statistics that provide valid, accurate and relevant measures of the economy and our society should be the starting point. The next step should be what available data and analysis of these data can offer. This requires getting an overview of already available data, in a data warehouse and data integration, including the combination of business and household data, which traditionally have been two isolated domains in official

statistics. It is only when we have a clear understanding of what data are missing that new new secondary sources can be investigated, and surveys can be designed.

In an integrated system of statistics the qualities of all data sources should be considered in order to get an optimal mix. All ways of data collection have their own strengths and weaknesses, as are summarized in table 2. For the usage of all data sources it is important to know about their conceptualization and operationalization, i.e. a metadata framework of the data, and how the data are collected. For administrative data, and especially for big data, this can be troublesome. In particular this is a challenge with regard to Big Data, as the objectives for these data are often unclear. Considerations for all sources with regard to official business statistics are stability and representativity.

As for primary data collection the challenge is tailoring the surveys to the business context. Surveys should no longer reflect the statistical system, but be tailored to the business information structures and production processes. This requires careful profiling of businesses, including not only contact information aimed at maintaining the relation, but also opening up the businesses.

We conclude with Bob Groves' words as stated at the 2013 NTTS conference: "We are living in exciting times: it is up to us to build a new paradigm for official statistics. We have work to do!"

Table 2: Basic quality issues of data sources.

	Surveys	Admin Data	Big data
Source of data	- Local	- Central	- Human sourced - Process mediated - Machine generated
Quality model	- Conceptual and operational: targeted and controlled	- Conceptual and partly operational	- Partly conceptualized, not operationalized
Qualities	- Designed - Multivariate	- Completeness - Low cost	- Timeliness - Real time updates
Disqualities	- Bias because of nonresponse	- Questionable validity - Not updated	- Bias because of coverage problems - Measurement errors - Lean in variables

References

Bethlehem, J., and Biffignandi, S. (2012), Handbook of Web Surveys, Wiley, Hoboken, NJ.

Buelens, B., Daas, P., Burger, J., Puts, M., and van den Brakel, J. (2014). Selectivity of Big data. Report Presented at the Statistics Netherlands Advisory Council meeting on 11 Feb 2014. Statistics Netherlands, Heerlen.

Burger, J., Davies, J., Lewis, D., van Delden, A., Daas, P., and Frost, J-M. (2013), Quality guidance for mixed-source statistics. Deliverable 6.3 of ESSnet Admin data, February.

Daas, P.J.H., van Delden, A. (2013), Collection and Use of Secondary Data. Module for the Handbook on Methodology for Modern Business Statistics (Memobust).

Daas, P.J.H., Ossen, S.J.L. (2011), Metadata Quality Evaluation of Secondary Data Sources. International Journal for Quality Research 5, 57–66.

Daas, P. J. H., Ossen, S.J.L., Tennekes, M., and Burger, J. (2012), Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

Daas, P.J.H., Puts, M.J.H. (2014), Big data as a Source of Statistical Information. The Survey Statistician 69, 22-31.

Dicken, P. (2007). Global Shift: Mapping the Chaning Contours of the World Economy. London: Sage.

Eurostat (2011), European Statistics Code of Practice, Eurostat, Luxembourg.

Fortanier, F., Korvorst, M., & Luppes, M. (2012). The employment Consequences of Globalization: linking data on employers and Employees in the Netherlands. In A. Dicaccio, M. Coli, & J. Ibanez, Advanced Statistical Methods for the Analysis of Large Data-Sets (pp. 101-111). Springer Verlag, Berlin/Heidelberg.

Frost, J. M. (2011), Development of Quality Indicators for Business Statistics Involving Administrative Data. Paper for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., and Khan, A. (2013), What does "Big Data" mean for Official Statistics? Paper for the High-Level Group for the Modernization of Statistical Production and Services, March 10.

Gravem, D., Löfgren, T., Haraldsen, G. and Bergstrøm, Y. (2011), Research on instrument design strategies to reduce response burden. In Giesen, D. and Raymond-Blaess, V. (eds): Inventory of published research: Response burden and reduction in official business statistics; A literary review of national statistical institutes' practices and experiences. Deliverable 2.1, Blue Ets project. European Commission Research Area, seventh framework programme.

Groves, R. M. (1989), Survey Errors and Survey Costs, Wiley, New York.

Groves, R.M. (2013), Official Statistics and "Big Data". Key note address at the 2013 European Conference for New Techniques and Technologies for Statistics (NTTS), 5-7 March 2013, Brussels

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004), Survey Methodology, Wiley, Hoboken, NJ.

Laitila, T, Wallgren, A., and Wallgren, B. (2011), Quality Assessment of Administrative Data. Research and Development – Methodology reports from Statistics Sweden, 2011:2, Stockholm/Örebro, Sweden.

Luppes, M., & De Winden, P. (2007). Open Economy and its relation to Employment and Welfare. Parijs: OECD.

OECD. (2013). Interconnected Economies: Benefiting from Global Value Chains (synthesis report). Paris: OECD.

Snijkers, G. (2008). Getting Data for Business Statistics: A Response Model. Proceedings of the 4th European Conference on Quality in Official Statistics, Rome, Italy.

Snijkers, G., (2009). Getting Data for (Business) Statistics: What's new? What's next? Paper presented at the 2009 European Conference for New Techniques and Technologies for Statistics (NTTS), 18-20 February 2009, Brussels.

Snijkers, G., Göttgens, R., and Hermans, H. (2011), Data collection and data sharing at Statistics Netherlands: Yesterday, today, tomorrow, paper presented at the 59th Plenary Session of the Conference of European Statisticians (CES), United Nations Economic Commission for Europe (UNECE), Geneva, June 14–16.

Snijkers, G., Haraldsen, G., Jones, J., & Willimack, D. (2013). Designing and Conducting Business Surveys. Hoboken, New Jersey: John Wiley & Sons, Inc. .

Statistics Norway (2007), Strategy for Data Collection, Statistics Norway, Oslo.

Sturgeon, T. (2013). Global Value Chains and Economic Globalization - Toward a New Measurement Framework. Boston: MIT.

UNECE (2007), Register-based statistics in Nordic countries – review of best practices with focus on population and social statistics. United Nations Publication, Geneva.

UNECE. (2011). The Impact of Globalization on National Accounts . Geneva: UN/ECE.

Weisberg, H. F. (2005), The Total Survey Error Approach: A Guide to the New Science of Survey Research, University of Chicago Press, Chicago.

Van der Veen, G. (2007). Integration of microdata from business surveys and the social statistics. DGINS Conference "The ESS response to globalisation - are we doing enough?". Budapest.

Winkler, W. (1995). Matching and record linkage. In B. e. Cox, Business Survey Methods (pp. 355-380). Wiley: New York.

Virtual Sprint paper (2014), How big is Big Data? Exploring the role of Big Data in Official Statistics. Paper of the first Virtual Sprint session for the Big Data HLG project (in preparation, available at: http://www1.unece.org/stat/platform/display/bigdata/How+big+is+Big+Data).

Zhang, L-C. (2012), Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66, 41–63.