Statistics Netherlands

System-to-System Data Collection in business surveys applied to an agricultural survey: small-scale pilot results [1]

Ger Snijkers, Tim de Jong, Chris Lam, and Cath van Meurs [2]
Statistics Netherlands, Heerlen/The Hague, contact: g.snijkers@cbs.nl

Paper presented at the 2024 UNECE Expert Meeting on Statistical Data Collection and Sources
22-24 May 2024, Geneva, Switzerland

1. Introduction

In the 20th century sample surveys have been the main data collection method for National Statistical Institutes (NSIs) to collect data (Snijkers et al. 2013). Even though secondary sources, like registers, have been used more and more since the 1970's, and more recently big data sources are being explored, at Statistics Netherlands (CBS) surveys still are the main data collection method (European Commission 2019; Snijkers et al. 2023a). Surveys are a primary data collection method, meaning that the data are collected directly from the sampled units. Sample surveys have proven to be a cost-efficient method to produce accurate statistics, although they come with a high cost both for the National Statistical Institutes (NSIs) and businesses, who may experience high response burden.

Completing (business) survey questionnaires involves a number of steps, which are described in the survey methodology literature (see e.g. Willimack and Snijkers 2013; Bavdaz 2007a). The core of these descriptive models consists of a 4-step question-and-answer model described by Tourangeau in the 1980's (1984; Tourangeau, Rips and Rasinski 2000; Snijkers 2002):

1. Comprehension of a question and the task(s) to answer the question
2. Retrieving the internal data needed to answer the question
3. Computation and evaluation of the answer
4. Reporting of the answer.

For business respondents steps 2 and 3 generally need a lot of work. For step 2, this includes collecting the data from internal data sources, which may involve colleagues from other departments. The retrieval process becomes more complicated in cases of mismatch between the requested and available data. This is followed in step 3 by calculating the required answers (often totals) based on the retrieved data. Studies show that these activities are considered burdensome (see e.g. Haraldsen 2018; Snijkers, Houben and Demollin 2023c). A record-keeping study conducted by CBS for the Crop Protection Survey questionnaire (Snijkers and Wieling 2020) showed that farmers have the required data (on event level) in their crop registration system (Farm Management Information System, FMIS). However, retrieving the data and calculating the required totals takes a lot of time. Consequently, farmers may adopt a satisficing response behavior (Krosnick and Presser 2010) resulting in measurement errors. Farmers asked for a less burdensome and more efficient way of reporting. With modern technologies, both the retrieval and computation step can be automated provided that the data are electronically available, and can be accessed using IT technologies.

Nowadays in the information age, there are a lot of new digital data sources in smart industries, also called "Industry 4.0" (see e.g. Haverkort and Zimmermann 2017), such as smart (or precision) farming. Charania and Li (2020), Pham and Stack (2018), and CEMA (2020) provide an overview of agricultural technologies applied in precision farming. They include machines, drones and robots with sensors, collecting a large number of data on crops, growing conditions and farming activities. Snijkers et al. (2021) discuss the options for precision farming as input for official statistics.

A large producer of precision farming machines is John Deere (Wikipedia 2023): their machines are equipped with a large number of sensors; the collected data are stored in a large relational database in the cloud: MyJohnDeere. Increasingly, these data sources provide Application Programming Interfaces

[1] This project is funded by Eurostat: EU grant 101036345-2020-NL-AGRI-SISA (Paulussen 2020).

[2] **Acknowledgement:** For their indispensable contributions we would like to thank: Bianca Aarts, Bryan Bungardt, Arthur Denneman, Renzo Ghianni, José Gómez-Pérez, Martin Klausmann (John Deere), Marius Reitsema, Ralph van Geenen, and Rob Vijftigschild.

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

(APIs)[3], through which these data are made available, as is the case for the MyJohnDeere cloud. This API-based communication process allows systems to share data without human intervention and makes System-to-System (S2S) data collection possible (Bharosa et al. 2015; Buiten et al. 2018; GBNED 2022). In theory, a S2S data communication method would reduce the response burden for farmers and businesses in general: data retrieval and computation of the requested answers would be done automatically. Instead of having farmers/businesses complete CBS questionnaires manually, this process can be automated, thus combining IT technology and survey methodology.

Based on this idea, we started a project to study the use of these sensor-based data in the completion of CBS agricultural surveys. In January 2020, we discussed this idea with a number of manufacturers of agricultural machines participating in the ATLAS project[4] (like Claas, New Holland, and John Deere; ATLAS 2020), after which John Deere came forward to collaborate with CBS.

Based on the MyJohnDeere system, a S2S data collection method was developed aimed at automatically completing an electronic CBS agricultural questionnaire: the Crop Yield Survey questionnaire. The goal of this project was the development of a proof-of-concept of such a S2S method for automatically collecting business data. This project was not aimed at implementing it in the Crop Yield Survey, although we worked towards this. The first steps for this project were taken in 2020, after John Deere had contacted us (and presented at EESW 2021: Góméz Pérez and Snijkers 2021). The project actually started in January 2022 as Work Package 1 (WP1) of the larger Eurostat Agri-Sisa project (Paulussen 2020; Snijkers et al. 2022a), and was ended in October 2023 (Snijkers et al. 2023b).

After this WP1 project was started, CBS adopted a new vision on primary business data collection in which businesses are put first (Nieuwenhuijs et al. 2022; Nieuwenhuijs, Houben, and Snijkers 2022; see also Bavdaz et al. 2020). This means that CBS has to move towards businesses and try to tailor their data collection designs to the business context as much as possible. A spot on the horizon specified in this vision is to maximize the use of existing data within businesses by automated retrieval of these data. This WP1 project fits within this vision, and serves as one of the first proofs of concept for a S2S approach.

2. The basic idea: the System-to-System data collection method

With the availability and use of data registration systems for farmers (and businesses in general) it becomes worthwhile to study the use of these business data sources in the questionnaire completion process. Instead of having farmers complete CBS questionnaires manually, the idea is that the data (they have available about their own business) are used to pre-fill the questionnaire. Figures 1 and 2 show how we envision this new process: Figure 1 shows the high-level completion process; figure 2 shows a more detailed flow chart of this process. The S2S IT architecture is discussed in more detail in Section 5.

The starting point is an electronic questionnaire (eQ), which for CBS are Blaise questionnaires. The questionnaire is adapted in such a way that it offers the option of pre-filling. For the S2S data collection a two-step microservice architecture has been developed, consisting of two microservices: an authentication and data collection microservice. The authentication microservice makes sure that the farmer can log-in to his MyJohnDeere data in the cloud. The data-collection microservice acts like an intermediary between the Blaise questionnaire and the John Deere cloud: it collects the required data, calculates the answers to the questions in the questionnaire, and communicates those to the questionnaire. The Blaise questionnaire can receive these answers, pre-fills the appropriate data boxes, so that they can be shown to the farmer.

The envisioned completion process for farmers starts by switching on their computer, and opening their web browser:

1) A sampled farmer logs onto the online Blaise questionnaire as usual (see presentation slide 5). Sampled units receive an invitation letter with login details, including a web address, user name and password. After having opened the web page and after having entered the username and password, the eQ opens immediately.
2) Instead of showing the regular questions, now the farmer first is asked whether they have John Deere, use MyJohnDeere, and if they would like to use their MyJohnDeere data to pre-fill the questionnaire (see slides 7 and 8).

[3] According to Red Hat (2017): "An API is a set of definitions and protocols for building and integrating application software. [...] APIs let your product or service communicate with other products and services without having to know how they're implemented".
[4] ATLAS is a European project aimed at building an agricultural interoperability and analysis architecture: www.atlas-h2020.eu/objectives/.

Figure 1. Microservice architecture for S2S data collection based on APIs.

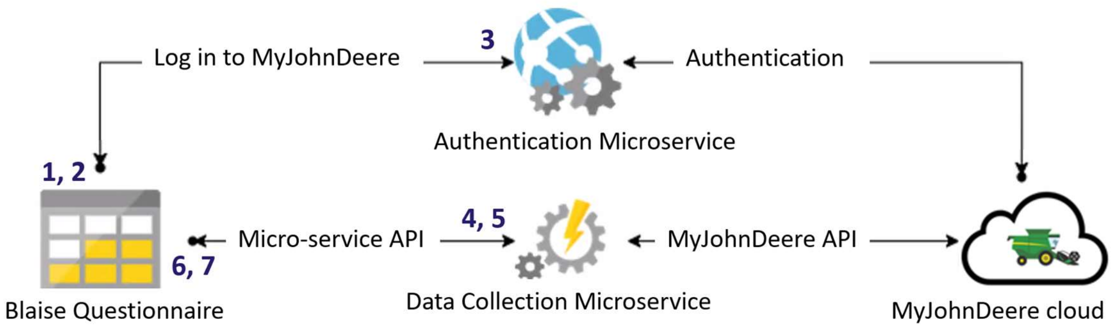
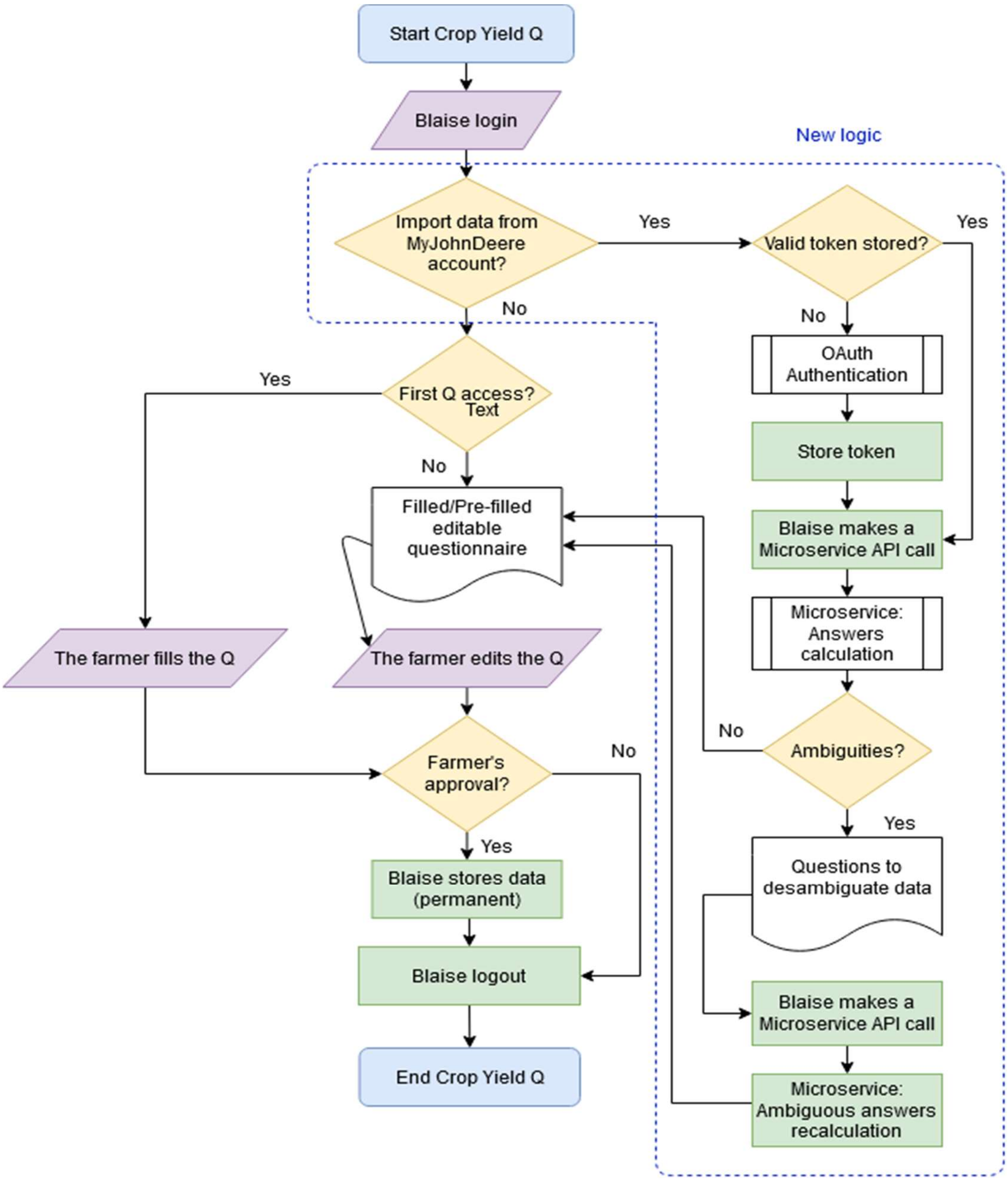


Figure 2. System-to-System Process sequence (flow chart)


Start Crop Yield Q
Blaise login
New logic
Import data from MyJohnDeere account?
Yes
Valid token stored?
Yes
No
No
OAuth Authentication
Yes
First Q access? Text
Store token
No
Filled/Pre-filled editable questionnaire
Blaise makes a Microservice API call
Microservice: Answers calculation
The farmer fills the Q
The farmer edits the Q
No
Ambiguities?
No
Farmer's approval?
Yes
Yes
Blaise stores data (permanent)
Questions to desambiguate data
Blaise logout
Blaise makes a Microservice API call
End Crop Yield Q
Microservice: Ambiguous answers recalculation

3) In case the answer is "yes", the farmer is redirected to MyJohnDeere (see slide 10 and 11), and follows an authentication process that gives CBS temporary and partial access to their MyJohnDeere data. This is done without sharing the farmer's credentials. (This authentication protocol is based on the standard delegation protocol OAuth 2.0). Now, the data collection microservice can make API calls to the John Deere cloud (step 4). In case the answer is "no", the farmer has to complete the questionnaire in the traditional way.
   (This is the envisioned, ideal process. Testing revealed that in practice authentication didn't work in this way: unfortunately a two-step authentication procedure was needed(see section 3).)
4) Via the "Microservice API" the electronics questionnaire asks for answers to the "Data-collection Microservice". The microservice browses the John Deere cloud looking for the appropriate data. These data are retrieved from the John Deere cloud via the "MyJohnDeere API", and kept in memory until the answers to the questions are calculated based on them. Immediately after, the answers are sent to the online questionnaire and pre-filled at the right locations.
5) In the context of the Crop Yield questionnaire, in some cases we find ambiguities related to the identification of crops, e.g. summer/winter crops per field. For instance, the crop harvested in a specific field is tagged as "wheat" without seeding data. In this case there is not enough information to classify it as "winter wheat" or "summer wheat". An extra step is introduced between steps 4 and 6 where the farmer is asked (through a web form) to select "winter wheat" or "summer wheat" for the ambiguous field. At this point, the online questionnaire makes a second API call, the microservice recomputes the ambiguous crop totals ("winter wheat" or "summer wheat" in this example) and sends the updated answers to the questionnaire. (This step is not implemented in the proof-of-concept.)
6) The pre-filled answers are presented to the farmer in the questionnaire (see slide 12). The farmer can check and edit the pre-filled questionnaire. Questions that could not be pre-filled still have to be completed manually.
7) After having checked and completed all questions, the farmer decides whether or not to send the answers. They can decide to start a next session at another time. The process ends when the answers to the questions are submitted (see slide 14).

3. Testing the system

The developed prototype was tested in three ways: the first test was prototype testing with John Deere test data during development); next a technical was done to test whether system technically works correct, and finally a small-scale field test with five farmers was conducted.

3.1. Prototype testing

The pilot prototype of the S2S architecture as discussed above, has successfully been tested in the "sandbox" mode, i.e. a playground for testing applications. The MyJohnDeere platform offers an ecosystem to run applications developed by third party software developers (that provide digital services to John Deere customers, which would include our S2S application). Applications can be run in two different modes: "sandbox" (only for testing purposes) and production. In order to run our system in the "sandbox" mode, we created a virtual farm in the platform that has been fed with open data provided by John Deere through its GitHub public repository. This allowed us to test the S2S communication in a way that is very close to real-life conditions. The "sandbox" test showed that technically our system works well: the data-collection microservice calculated the answers to the questions in the Crop Yield Survey questionnaire. The next step is a test under real-life conditions: a small-scale field test. This required the implementation of the system within the CBS environment.

3.2. Technical test

Before using the S2S approach in the field, the implemented system needed to be tested in a dry run. It is always required to do a 'logistics or technical test' to see if all the environment parameters of the new system are set right, before using it on a large scale in the field. Here we can see if the correct questionnaire is being used, if the microservices are reachable and if the URL-callback is functional within the IT landscape (in production or test server in our case). We concluded that the system works in theory, which is already a conclusion in itself: the technology works! The next step was to test it in the field.

3.3. Small-scale testing

Once the S2S architecture had been successfully tested, we conducted a small-scale field test to test the proof-of concept in real life situations. Such a field test, while small and qualitative in nature, is an important source of information regarding the operationalisation of the system, i.e. the way the S2S architecture will be presented to users. A small-scale test helps identify e.g. usability issues and provides us with valuable insights into the opinions of farmers about the chosen operationalisation. Also, it provides directions for improvements before a large-scale implementation in a field survey. In questionnaire design, the pre-testing of a draft of the questionnaire before using it in the field, is common practice (see e.g. Snijkers 2002).

The goal of the field test was twofold: (1) to study how the system itself works in practice (technical issues), and (2) to study how farmers perceive this approach (usability issues, trust). To this end, a test protocol was developed (see Appendix 3). In the spring of 2023, five farmers in various parts of the Netherlands were visited for a test interview following the test protocol. This Section discusses the set-up (recruitment of farmers and the interviews) and the results of the field test.

A number bugs and issues regarding the system unfolded as the interviews were conducted. The results from the test interviews can be clustered into five groups:

1. Technical and organizational issues: the field test identified several challenges from different perspectives. The main issue was that farmers had to authenticate prior to logging in onto the questionnaire. Authentication thus was a two-step procedure, which turned out to be cumbersome for farmers. Also the system as a whole was instable at times.
2. The farmer's perspective, like usability issues with the proposed S2S approach, questionnaire design issues, and fundamental objections of farmers to data sharing. It turned out that the presentation of the retrieved data was not immediately clear to farmers.
3. Data quality issues, which included sensor calibration issues, missing data (because of use of other brands and John Deere machines not connect to MyJohnDeere; contractors), unit issues (data from other farmers), and the small market share of John Deere. Also it turned out that farmers used the data in their Farm Management Information System (FMIS) to check the data. Farmers indicated that for administrative purposes they use such a FMIS, like Dacom and AgroVision (the two largest tools in the Netherlands). They recommended to connect to one of these systems instead of MyJohnDeere, which was not designed as a management tool.
4. On a different, more fundamental level, the interviews showed reservations by farmers to share data with government agencies. This concerns issues related to trust and workload. Farmers indicate that this approach would not help them in completing the questionnaire: "This doesn't reduce the time I need as compared to filling in a questionnaire in the usual way."
5. Despite these issues and reservations, on the whole the farmers were positive about this S2S approach. They assumed that in the future it would indeed reducing their workload/response burden. Here, we should keep in mind that we talked a very selective group of farmers: those who are open to innovations.

4. Conclusions

The test showed that the technology worked: a working proof of concept is what we initially aimed for. In addition, we also hoped to implement this approach in the next Crop Yield Survey. Because of the interview results, we decided not to implement. We concluded that the operationalisation of this system was not mature enough, and needed substantial improvements: the risks of failure in the field were considered too high. Especially the concluding statement from farmers that this system would not save time, and that they rather would complete the questionnaire manually, was very relevant for us. More time and resources were needed to improve it, which we didn't have before the preparations of the fieldwork for the survey started.

Despite the issues as mentioned above, which made the farmers actively wonder about the relevance and acceptance among farmers of the tested S2S approach, in the end they positively evaluated this approach. They expressed their interest in and recognized the added value of this approach, provided that it works well with regard to all issues discussed above: the methodology in itself is worthwhile, the operationalisation needs improvements.

Here, we should keep in mind that we talked to a group of innovative farmers. They most certainly have specific views on these developments which cannot be generalized to farmers as a whole. On the other hand, when implementing this approach, we need to target these innovative farmers first: they are the early adopters. These farmers are open minded and are intrinsically curious towards leveraging farm data to a higher level, in order to enhance their agricultural practices and improve their businesses. The system-to-system approach aligns well with their visions. Moreover, farmers provided suggestions for better data sources to connect to: their FIMS systems Dacom and Agrovision. Another Agri-Sisa Work Package (WP3, Paulussen 2020, Snijkers et al. 2022b; Denneman et al. 2023) explored these systems, and resulted in a positive business case. Our next steps will be working towards implementing an S2S application with these FMIS systems.

References

ATLAS (2020), ATLAS Newsletter, issue #1, April 2020. (Available at: https://www.atlas-h2020.eu/).

Bavdaž, M. (2007a), Measurement Errors and the Response Process in Business Surveys, PhD thesis, University of Ljubljana, Slovenia.

Bavdaž, M., G. Snijkers, J.W. Sakshaug, T. Brand, G. Haraldsen, B. Kurban, P. Saraiva, and D.K. Willimack (2020), Business data collection methodology: Current state and future outlook. *Statistical Journal of the IAOS*, *36*(3): 741-756.

Bharosa N., R. van Wijk, N. de Winne, and M. Janssen, eds. (2015), *Challenging the Chain. Governing the automated exchange and processing of business information*. IOS Press, Amsterdam, the Netherlands. (Available at: www.iospress.nl/book/challenging-the-chain/).

Buiten, G., G. Snijkers, P. Saraiva, J. Erikson, A.-G. Erikson, and A. Born (2018), Business data collection: Toward Electronic Data Interchange. Experiences in Portugal, Canada, Sweden, and the Netherlands with EDI. Journal of Official Statistics 34(2): 419-443 (ICES-5 special issue).

CEMA (2020), Precision farming: What is Precision farming all about? European Agricultural Machinery Association. Brussels, Belgium. (Available at: https://cema-agri.org/index.php?option=com_content&view=article&id=50: precision-farming&catid=10&Itemid=170 (accessed 7 August 2022)).

Charania I, and X. Li (2020), Smart farming: Agriculture's shift from a labor intensive to technology native industry. Internet of things, Vol. 9. (Doi: 10.1016/j.iot.2019.100142).

European Commission (2019), ESS (European Statistical System) Workshop on the use of administrative data for business, agriculture and fisheries statistics, 17-18 October 2019, Bucharest, Rumania**.** (https://ec.europa.eu/eurostat/cros/content/ess-workshop-use-administrative-data-business-agriculture-and-fisheries-statistics_en).

Denneman, A., R. Vijftigschild, L. Hoogervorst, C. van Meurs, M. Reitsema, R. Ghianni, G. Snijkers, T. de Jong, Ch. Lam, J. Gómez-Pérez, J. Backman, A. Ronkainen,and M. Yli-Heikkilä (2023), Farm Management Information Systems: A potential data source for agricultural statistics. AGRI-SISA: WP3 final report. Statistics Netherlands, The Hague/Heerlen, The Netherlands, and Natural Resource Institute Finland (Luke), Finland.

GBNED (2022), Accounting software: standards and interfaces (in Dutch: Boekhoudsoftware: standaarden en koppelingen). GBNED research bureau. (Available at: www.softwarepakketten.nl/bericht/7503&bronw=1/Boekhoudsoftware_standaarden_en_koppelingen.htm)

Gómez Pérez, J., and G. Snijkers (2021), A new method for automated business data colcleetion of official statistics through APIs. Paper presented at the 2021 European Establishment Statistics Workshop (EESW21), 14-17 September 2021, on-line Workshop, hosted by Statistics Netherlands, The Hague.

Haraldsen, G. (2018), Response processes and response quality in business surveys. In: Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-Ch., and Zimmerman, Th., eds., *The Unit Problem and Other Current Topics in Business Survey Methodology*: pp. 157-177. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.

Haverkort, B.R., and A. Zimmermann (2017), Smart industry: How ICT will change the game! IEEE Internet Computing: 21(1): 8-10. (Doi: 10.1109/MIC.2017.22)

Krosnick, J.A., and Presser, S. (2010), Questions and Questionnaire Design. In: Marsden, P.V., and Wright, J.D., eds., *Handbook of Survey Research*, 2nd Edition: pp. 263-313. Emerald Group Publishing Limited, Bingley, UK.

Nieuwenhuijs, R., et al. (2022), A new strategy on primary business data collection: "naturally relevant" (in Dutch: Toekomstvisie primaire waarneming bedrijven: CBS - vanzelfsprekend relevant). Statistics Netherlands, The Hague/Heerlen.

Nieuwenhuijs, R., L. Houben, and G. Snijkers (2022), A new vision on primary data collection from businesses: "naturally relevant". Paper presented at the 2022 UNECE Expert meeting on Statistical data Collection, Rome, 26-28 October 2022. (Available at: https://unece.org/statistics/events/DC2022).

Paulussen, R. (2020), Project description for 2020-NL-AGRI-SISA: Agriculture system integration and spatial analysis. Project Technical Description regarding Modernisation of Agricultural Statistics: ESTAT-2020-PA8-E-AGRI. Statistics Netherlands (CBS), The Hague, The Netherlands, and Natural Resource Institute Finland (Luke), Finland.

Pham, X., and M. Stack (2018), How data analytics is transforming agriculture. Business Horizons, Vol. 61: 125-133. (Doi: 10.1016/j.bushor.2017.09.011).

Snijkers, G. (2002), *Cognitive Laboratory Experiences: On Pre-Testing Computerized Questionnaires and Data Quality.* PhD thesis Utrecht University. Statistics Netherlands, Heerlen.

Snijkers, G., M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, K.J. Thompson, and A. van Delden, eds. (2023a), *Advances in Business Statistics, Methods and Data Collection*. Wiley, Hoboken.

Snijkers, G., T. de Jong, and B. Bungardt (2022a), System-to-system data communication applied to the CBS Crop Yield Survey using John Deere. AGRI-SISA: WP1 intermediate report. Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Snijkers, G., T. de Jong, Ch. Lam, C. van Meurs, B. Bungardt (2023b), System-to-system data communication applied to the Statistics Netherlands Crop Yield Survey using MyJohnDeere. AGRI-SISA: WP1 final report. Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack, eds. (2013), *Designing and Conducting Business Surveys*. Wiley, Hoboken.

Snijkers, G., L. Houben, and F. Demollin (2023c), Tailoring the design of a new combined business survey: process, methods, and lessons learned. In: Snijkers, G., M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, K.J. Thompson, and A. van Delden (eds.), *Advances in Business Statistics, Methods and Data Collection*: pp. 357-386. Wiley, Hoboken.

Snijkers, G., T. Punt, S. De Broe, and J. Gómez Pérez (2021), Exploring sensor data for agricultural statistics: The fruit is not hanging as low as we thought. Statistical Journal of the IAOS 37(4): 1301-1314. (DOI 10.3233/SJI-200728).

Snijkers, G., R. Vijftigschild, T. de Jong, J. Gómez-Pérez, A. Denneman, J. Backman, A. Ronkainen, and M. Yli-Heikkilä (2022b), Farm Management Information Systems: A potential data source for agricultural statistics. AGRI-SISA: WP3 intermediate report. Statistics Netherlands, The Hague/Heerlen, The Netherlands, and Natural Resource Institute Finland (Luke), Finland.

Snijkers G, and M. Wieling (2020), Pre-testing the 2020 crop protection survey questionnaire: Set-up and results (in Dutch: Pre-test vragenlijst Gewasbescherming 2020: Opzet en resultaten). Internal report. Heerlen (The Netherlands): Statistics Netherlands, Methodology Department.

Tourangeau, R. (1984), Cognitive science and survey methods. In: Jabine, T., Straf, M., Tanur, J. M., and Tourangeau, R., eds., *Cognitive Aspects of Survey Design. Building a Bridge between Disciplines*: 73-100. National Academy Press, Washington DC, USA.

Tourangeau, R., L.J. Rips, and K. Rasinski (2000), *The Psychology of Survey Response*. Cambridge University Press, New York.

Wikipedia (2023), John Deere. https://en.wikipedia.org/wiki/John_Deere (last checked: 3 May 2024).

Willimack, D.K. (2013), Methods for the Development, Testing, and Evaluation of Data Collection Instruments. In: Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack, eds., *Designing and Conducting Business Surveys*: pp. 253-301. Wiley, Hoboken.

Willimack, D.K., and G. Snijkers (2013), The business context and its implications for the survey response process. In: Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack, eds., *Designing and Conducting Business Surveys*: pp. 39-82. Wiley, Hoboken.