UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Expert Meeting on Statistical Data Collection – Towards a New Normal?
26 to 28 October 2022, Rome, Italy

Session 2
19 September 2022

System-to-System Data Collection in business surveys applied to an agricultural survey: a Proof of Concept

Ger Snijkers; José Gómez Pérez; Tim de Jong (Statistics Netherlands)
g.snijkers@cbs.nl; tja.dejong@cbs.nl

Abstract

At the end of the 2018 edition of this workshop, I pitched the idea of automated data collection for business surveys. In 2019, I presented the results of a first exploratory study. Now, 3 years later we have a Proof of Concept.

In the 20th century, 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. Nowadays in the information age, there are a lot of new digital data sources in smart industries, like in precision farming. In some cases, these data sources allow for data communication with other computer systems without human intervention via Application Programming Interfaces (APIs). Based on these software interfaces, we developed a system-to-system data collection methodology that reduces response burden by automating the business's internal data retrieval process. Applied to the official Crop Yield Survey, a software prototype was developed based on this methodology.

At the workshop, we will present the IT architecture we developed, showing how data capture and processing can be automated. We will discuss the automated pre-filling of the electronic Crop Yield Survey questionnaire using an API provided by a smart farming machine manufacturer, John Deere: the MyJohnDeere API. In a first Proof of Concept, it has been applied to data from a virtual farm, showing that it works, and the farmer's workload to complete a questionnaire can be limited to a minimum.

Our next step is to conduct a small-scale field test with a small number of farmers to study the method in practice. This field test is planned for the fall of 2022. Hopefully, we can present the first results of this field test at this Expert Meeting in October.

It is our belief that this system-to-system method can be applied to business surveys in general and in the future will replace the manual completion of business survey questionnaires including the manual retrieval and re-keying of data.

Statistics Netherlands

System-to-System Data Collection in business surveys applied to an agricultural survey: a Proof of Concept

Ger Snijkers, José Gómez Pérez, and Tim de Jong
Statistics Netherlands, Methodology Department, Heerlen
Contact: G. Snijkers, g.snijkers@cbs.nl

Paper presented at the 2022 UNECE Expert Meeting on Statistical Data Collection 'Towards a New Normal?', 26-28 October 2022, ISTAT, Rome.

1. Introduction

Sample surveys are a primary data collection method. In the 20th century 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 (Snijkers et al., 2013). 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 (see e.g. Pham and Stack, 2018; Snijkers et al., 2021). Increasingly these data sources provide Application Programming Interfaces (APIs)[1], through which these types of data are available. This API-based communication allows systems to communicate without human intervention and makes System-to-System (S2S) data collection possible (Bharosa et al., 2015; Buiten et al., 2018).

An example of S2S that we will discuss in this paper involves the automated pre-filling of an electronic agricultural questionnaire using an API provided by a smart farming machine manufacturer, John Deere: the MyJohnDeere API. This could replace the manual completion of questionnaires including the manual retrieval and re-keying of data.

S2S data collection methods could allow NSIs to (Punt and Snijkers, 2019; Snijkers and Gómez Pérez, 2020):

a) Replace (parts of) surveys, in particular replace manual completion of survey questionnaires.
b) This would reduce data collection costs of NSIs (in the long-term) and businesses (i.e., response burden).
c) Develop new statistics (including real-time statistics) from the new data sources, and provide useful and more detailed information back to the businesses (closing the data cycle).

In this paper we will discuss a pilot prototype we developed to address the first goal. The prototype implements a S2S data collection method using APIs: data available in the MyJohnDeere cloud are automatically collected to pre-fill the Crop Yield Survey questionnaire. This questionnaire is sent yearly to sampled farmers to be completed at the end of each year. An exploratory study that we did prior to developing this prototype showed that the data required to complete this questionnaire are available in MyJohnDeere. These data in

[1] 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".

MyJohnDeere are generated by sensors in John Deere machines used in precision farming. In this study John Deere is used as a first Proof of Concept.

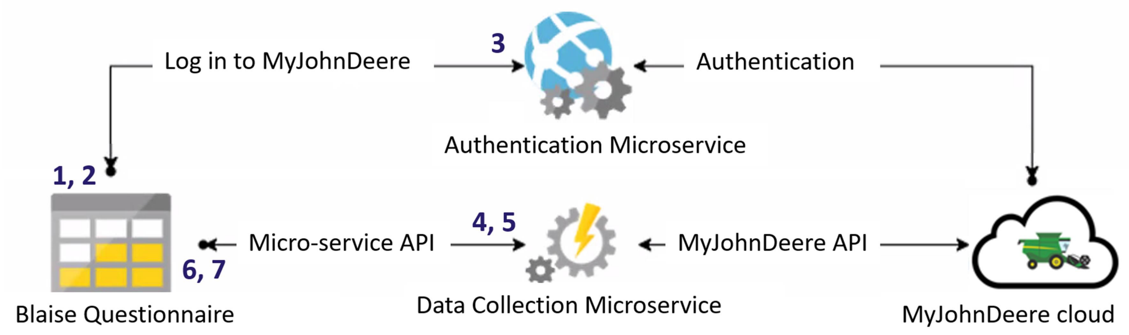
Even though this API-based method is developed for collecting agricultural data, we feel that this S2S data collection methodology can be applied generally for all business sectors as long as they match the criteria described in Section 4. In this paper, we discuss how we actually can collect these sensor data from businesses: we discuss a new method to collect sensor data through APIs, and apply this method to agricultural data.

This S2S methodology will be discussed in Section 2 of this paper. While we have tested the pilot prototype with open data (see Section 3), we plan to test the field pilot with real data from farmers (see Section 4). Section 4 concludes this paper.

2. Method

For the S2S data collection, we have chosen a microservice architecture, including an authentication and data collection microservice, as is shown in Figure 1. The authentication microservice makes sure that the farmer can log-in to his MyJohnDeere data in the cloud. Next, the data collection microservice acts like an intermediary between the electronic questionnaire (eQ, i.e. a Blaise questionnaire at Statistics Netherlands) and the John Deere cloud.

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



The process sequence farmers will follow, starts by logging onto the Blaise eQ:

1) A sampled farmer logs onto the online questionnaire as usual. In Statistics Netherlands (CBS) all questionnaires are electronically available. 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 keyed in the username and password, they enter the eQ immediately.
2) Instead of starting to complete the regular questions, now the farmer first is asked if they have John Deere smart farming machines, use MyJohnDeere and if they are willing to use their precision farming data in order to pre-fill the questionnaire.
3) In case the answer is “yes”, the farmer follows an authentication process that gives us temporary and partial access to their data in MyJohnDeere. This is done without sharing the farmer’s credentials. This authentication protocol is based on the standard delegation protocol OAuth 2.0. Now, the 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.

4) Via the “Microservice API” the online questionnaire asks for answers to the “Pilot 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. Right after, the answers are sent to the online questionnaire and imputed.
5) In the context of the crop yield questionnaire, in some cases we find ambiguities related to the identification of summer/winter crops per field. For instance, the crop harvested in a specific field is tagged as “wheat” and there is no 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 the 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 (“winter wheat” or “summer wheat” in this example) totals and sends the updated answers to the questionnaire.
6) The imputed answers are presented to the farmer in the questionnaire. The farmer can check and edit the pre-filled questionnaire. Questions that could not be pre-filled still have to be completed manually (if applicable).
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.

Using this architecture, eQ software functionality can be extended without modifying the questionnaire software itself as long as it supports the following two features:

a) Communication with other computer systems without human intervention via APIs,
b) Electronic handling of data access permissions through an authentication protocol.

Since these features are based on standard software practices, this methodology could be easily adopted by other NSIs.

Most of the S2S data collection process is automated. This methodology automates the completion of the Crop Yield questionnaire but it doesn’t change the questionnaire itself. During login and authentication, human intervention is always required and may also be needed during disambiguation and final editing of the questionnaire.

3. Prototype testing with open data and next step

Based on the MyJohnDeere platform, there is an ecosystem of applications developed by third party software companies that provides digital services to John Deere customers. Applications can be run in two different modes: "sandbox" (only for testing purposes; John Deere, 2022) and production. Our pilot prototype has been already successfully tested in the "sandbox" mode. In order to do that, we have created a virtual farm in the platform that have been fed with open data provided by John Deere through its GitHub public repository (John Deere, 2020).

The "sandbox" mode allowed us to test the S2S communication in a way that is very close to real conditions. This "sandbox" test showed that technically our system works well: the data collection microservice (as shown in Figure 1) calculated the answers to the questions in the crop yield survey.

The next step is a small-scale field pilot which we are preparing at this very moment. This field pilot the focus is on user experience (and response burden), costs, and data quality, in order to improve the architecture.

4. Going beyond agriculture, and beyond the questionnaire

Even though we still have to assess how this method will work in practice with data from farmers, this "sandbox" prototype demonstrates how data capture and processing can be automated.

There are a number criteria that need to be fulfilled for this methodology to be applied in general:

a) Significant overlap between the data source and the questionnaire
b) The data source must have an API for S2S communication.
c) For privacy reasons, electronic handling of data access permissions should be possible.

John Deere and other big farming machine manufacturers that have similar APIs (for instance CNH Industrial, New holland, and Claas) have international presence in markets all over the world. For this reason, our methodology can be applied to arable farming in other countries. In addition, we are studying Farm Management Information Systems (FMIS), the farmer's crop registration systems, in order to apply this methodology.

With regard to the Crop Yield Survey we focused on field operations data, but the John Deere cloud stores other types of data, like machine, agronomic service providers activity, soil and environmental conditions data. It is a huge potential data source for modernizing agricultural statistics (Wirtz, 2019).

Other agricultural surveys like the crop protection survey, and surveys for other statistics (for instance, transportation and finance are fields where we can find a lot of APIs) can benefit from this methodology. Nowadays in the information age, there are a lot of new digital data sources in smart industries, like in smart (or precision) farming. In some cases, these data sources provide APIs. One of our next steps is checking business sectors for these conditions.

In addition, for large scale usage, there are still a few data challenges that need to be fulfilled (Snijkers et al, 2021):

a) Data harmonization. The same methodology can be applied to other manufacturers clouds but the software would require partial rewriting.
b) Standardization of S2S. In order to minimize the rewriting of software as much as possible, standardization of S2S software is needed.
c) Stability of (meta)data architecture in the future.
d) Precision farming market penetration. Precision farmers are still a minority in the Dutch farming sector but at the same time they are bigger farms (van Merrienboer and Bakker-Smit, 2020; Snijkers et al, 2021).

We believe that this type of S2S data communication solutions is valuable and promising, and can go beyond the questionnaire, by leaving the questionnaire out of the process, but there is still a long way to go.

Acknowledgement

The authors are grateful to John Deere & Company. The work presented in this paper is possible thanks to its MyJohnDeere API and the Information Technology (IT) services around it.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the official policy of Statistics Netherlands.

References

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.

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).

Haverkort, B.R., and A. Zimmermann (2017), "mart Industry: How ICT Will Change the Game! *IEEE Internet Computing,* 21(1): 8-10.

John Deere (2020), SampleData. Available at: https://github.com/JohnDeere/SampleData.

John Deere (2022), Sandbox. Available in the Glossary: https://developer-portal.deere.com/#/myjohndeere/glossary

Pham, X., and M. Stack (2018), How data analytics is transforming agriculture. *Business Horizons,* 61: pp. 125-133.

Punt, T., and G. Snijkers (2019), Exploring precion farming data: a valuable new data source? A first exploration. Paper presented at the 2019 UNECE Workshop on Statistical Data Collection 'New Sources and New Technologies', 14-16 October 2019. Geneva (Switzerland). Available at: https://unece.org/statistics/events/DC2019

Red Hat, 2017, What is an API? Red Hat Inc., available at: https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces (accessed 21 May 2022).

Snijkers, G., and J. Gómez Pérez (2020), Exploring precision farming data: a valuable new source for official statistics? A pilot with System-to-System data communication applied to John Deere data. Presentation at the BigSurv20 conference (on-line), 13 November 2020, hosted by Utrecht University, Utrecht, Netherlands. Available at https://www.youtube.com/watch?v=jsPqHaCqhMo.

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

Snijkers, G., Punt, T., De Broe, S., 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.

van Merrienboer S, and G. Bakker-Smit (2020), Precision agriculture in the Dutch agricultural sector (in Dutch: Precisielandbouw in de Nederlandse akkerbouwsector). RaboResearch Food & Agribusiness, Rabobank, Netherlands. Available at: https://research.rabobank.com/far/en/sectors/regional-food-agri/precisielandbouw-in-de-nederlandse-akkerbouwsector%20.html.

Wirtz, Ch. (2019), The role of Eurostat in modernizing agricultural statistics. Paper presented at the 62nd ISI World Statistics Congress (STS-423, 19 August), 18-23 August 2019, Kuala Lumpur, Malaysia. Available at: https://www.isi2019.org/scientific-programme-2/.