**Smart business statistics:  
how to integrate technology and   
official statistics**

Ger Snijkers, Statistics Netherlands, [g.snijkers@cbs.nl](mailto:g.snijkers@cbs.nl) [[1]](#footnote-1)

Sofie De Broe, Statistics Netherlands, [smmg.debroe@cbs.nl](mailto:smmg.debroe@cbs.nl) 1

**Abstract**

*Businesses are requested to deliver a lot of data for official statistical purposes. Traditionally, questionnaires are used as collection instrument, which represents response burden for businesses and high costs for NSI and businesses alike.*

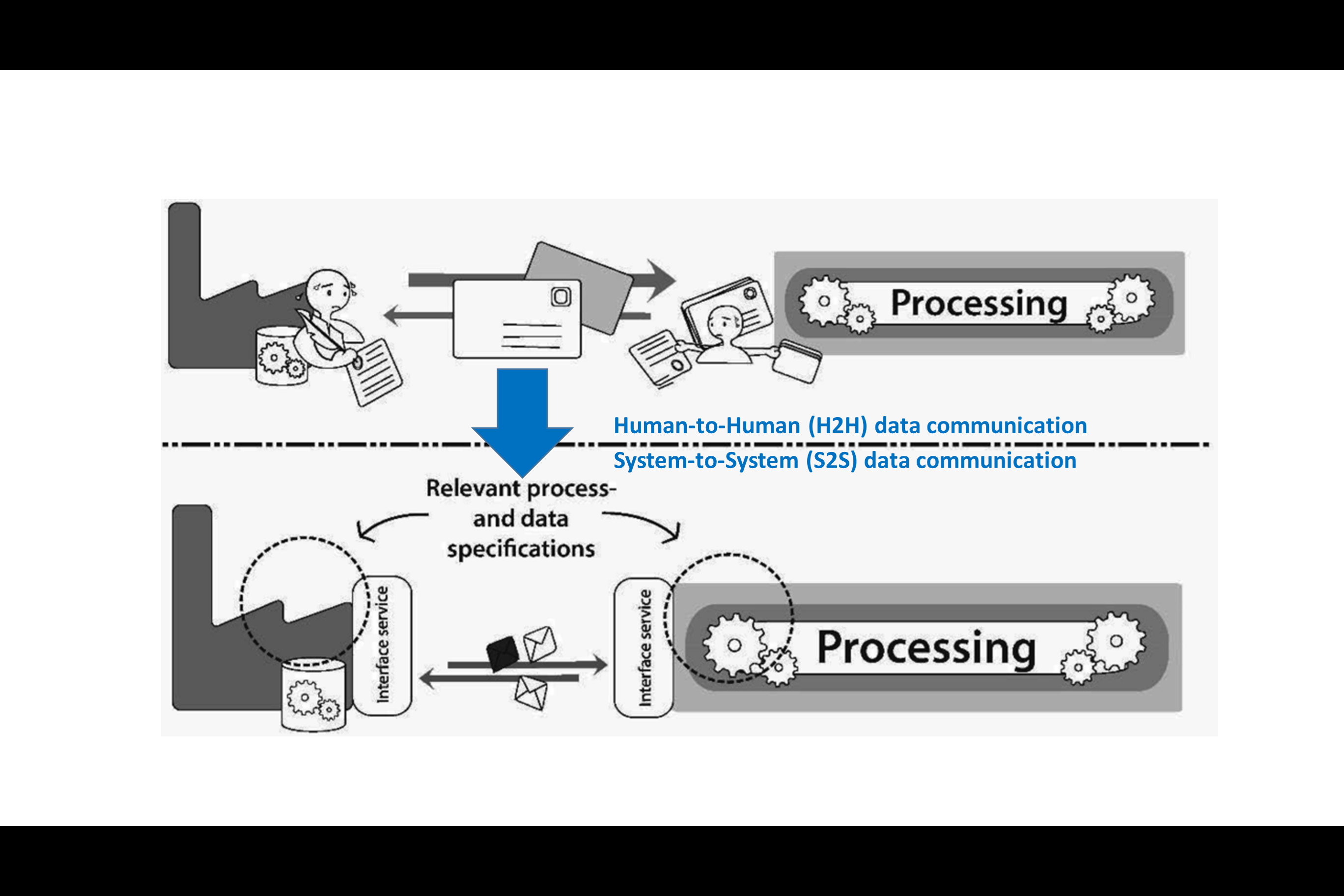
*Nowadays, however, a lot of data within businesses are already available in electronic format. Electronically available data could be collected by means of System-to-System (S2S) data-communication technology as a primary data-collection mode. A first example regarding S2S data collection for statistical purposes involves sensor data. Increasingly, electronic sensors are used to run a business, e.g. by agricultural businesses (like dairy farms using milking robots). A second example is financial data in a fully integrated and digitalised business information chain which makes S2S data communication for financial, tax and statistical reports possible.*

*In this paper we explore the enablers and challenges of S2S data collection in the context of business statistics. Drivers for switching from questionnaires to S2S are: working towards smart business statistics, (i.e. timely and new statistical output integrated in business processes), the reduction of response burden, and monitoring and benchmarking businesses to their counterparts. S2S data collection seems straightforward. However, important factors that affect or impede implementation include among others harmonisation of (meta)data, data quality issues, data-stewardship, and trust in data. S2S data collection offers opportunities to get access to data in a cost-effective manner (as opposed to sending out questionnaires); however, both businesses and NSIs may need to do initial investments to make the technology work, as well as relation building. In terms of getting access to the data and business participation, the question from businesses remains: “what’s in it for me?” Therefore, our statement in this paper is: Technology is the enabler of innovations; it is the applied methodology and organisational context that make innovations work.*

**Keywords:** automated data exchange, electronic data interchange (EDI), sensor data, data quality, enablers and challenges

**1. Introduction**

In the field of primary business data collection we foresee substantial changes: we will be moving from business surveys using web-based questionnaires to automated data exchange, also known as Electronic Data Interchange (EDI). Bharosa et al. (2015: 6) define EDI, as “the movement of business documents electronically between or within firms (including their agents or intermediaries) in a structured, machine-retrievable data format that permits data to be transferred, without re-keying, from a business application in one location to a business application in another location.” If we replace the word “business documents” by “business data”, EDI refers to data communication. As such we could say that we are moving from Human-to-System (H2S, i.e. web questionnaires) to System-to-System (S2S) data communication (Bharosa et al., 2015) between businesses and institutes like NSIs, other governmental agencies like Tax Offices, Central Banks, and Chambers of Commerce, as well as commercial banks, as is shown in figure 1. And if we think of “business data” not only of traditionally collected data, but as all kinds of data that are generated in business processes by smart sensing machines/robots which are connected via the Internet of Things for data communication, we could say that we are on the verge of the biggest change in official business statistics since the introduction of survey sampling in the early 1900’s. We will use S2S and EDI interchangeably.

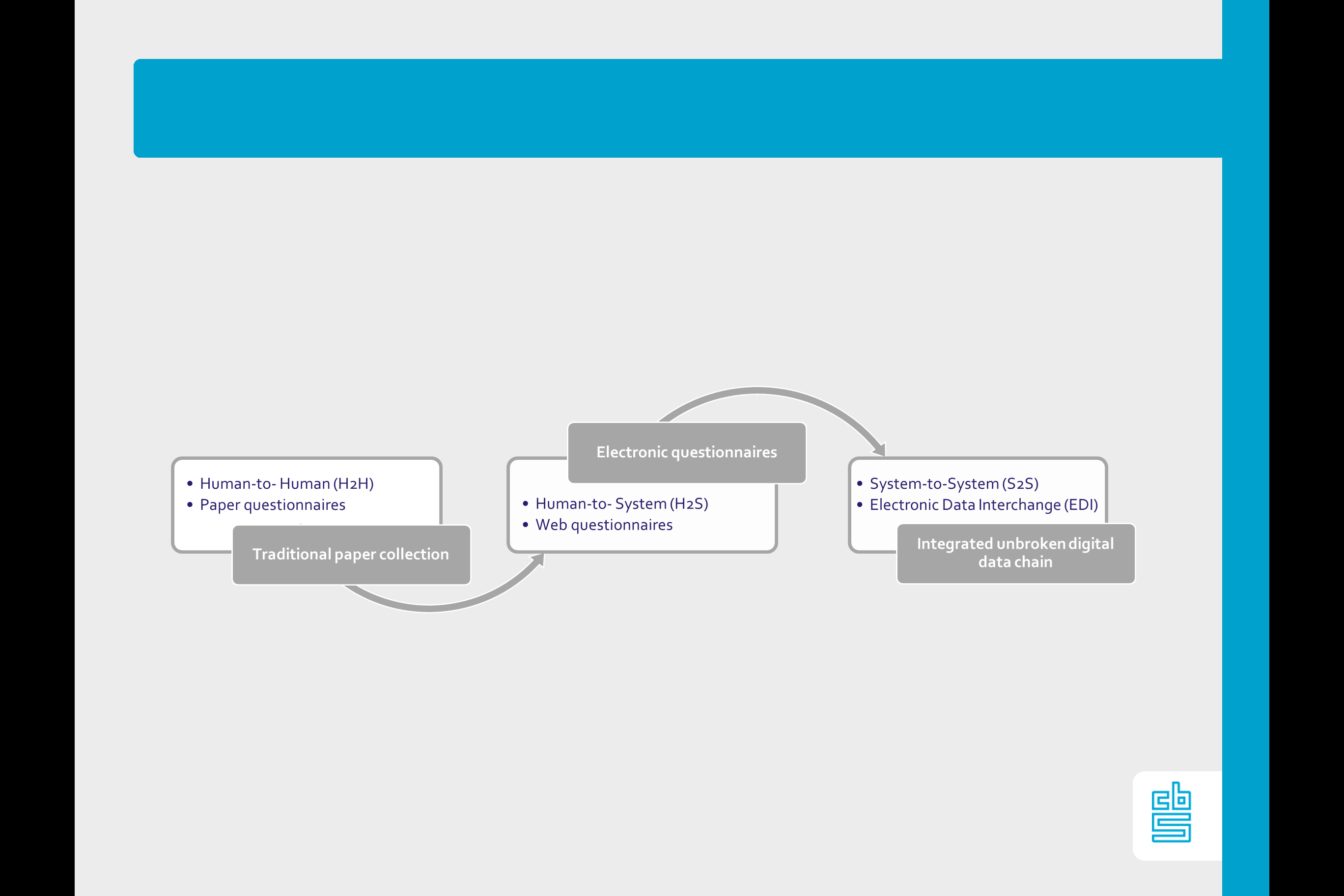
**Figure 1. From H2H to S2S** (adapted from Bharosa, 2015: figure 1.3)

In this paper we will discuss innovations in businesses making the application of EDI technology and S2S data collection approaches for business statistics possible. Although the technology for S2S data communication and EDI is not new, as we will see in section 2, its potentials have not been exploited to the full. Nowadays innovative developments in the business world, as well as developments solving methodological, organisational, legal and ethical issues concerning the application of EDI, make the application of this technique promising.

In section 3 we will discuss a number of innovative developments in the business world: these developments are the enablers for EDI. In Section 4 we will see that to make EDI work, a number of challenges need to be tackled. Section 5 concludes this paper with discussing smart statistics options. Referring to Srinivasan (2017) and Hinssen (2017) our statement is: Technology is the enabler of innovations; it is the applied methodology and organisational context that make innovations work.

**2. A brief overview of business data collection**

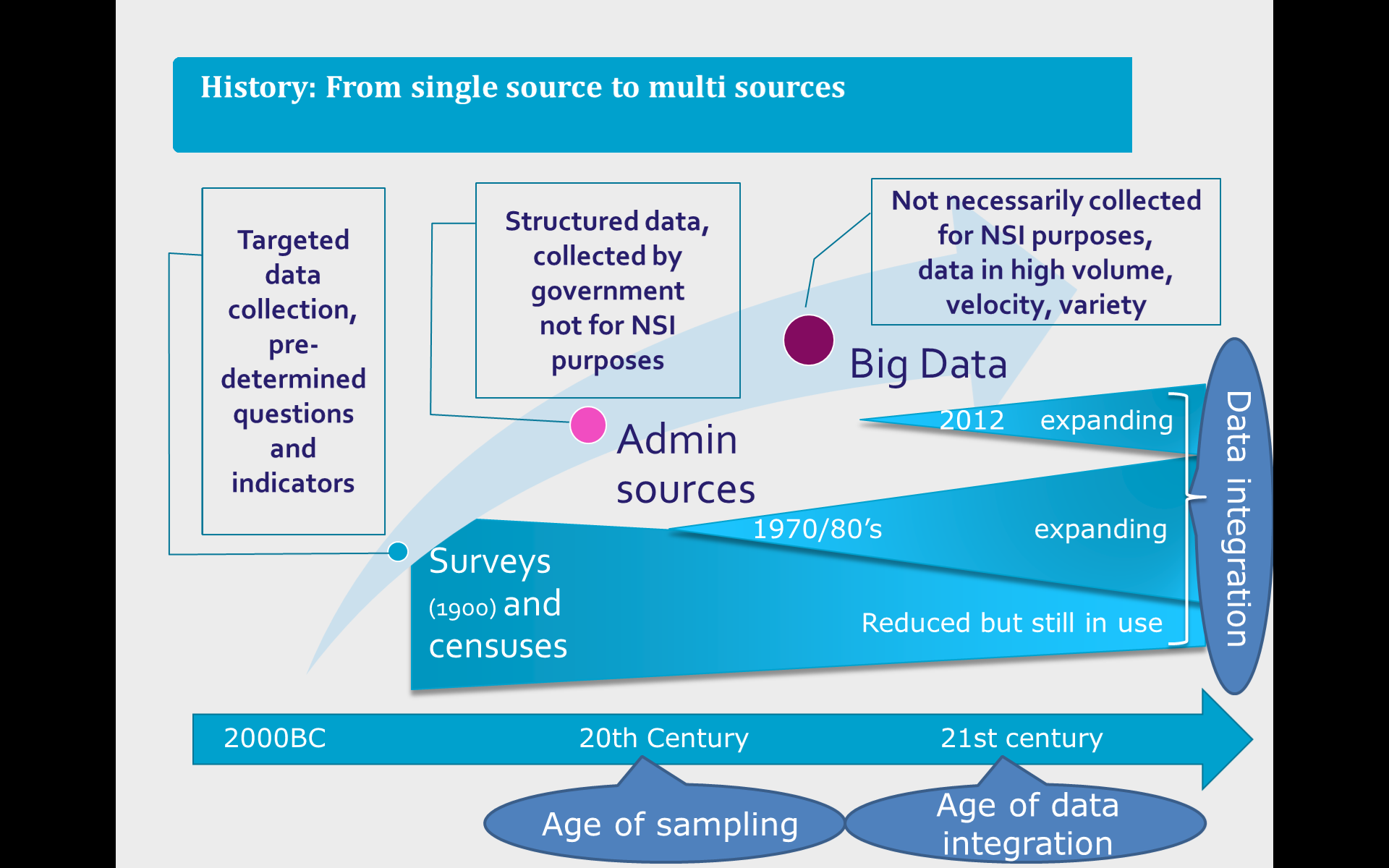
The development of moving from H2S to S2S primary data communication in the field of business data collection, already started at the end of the 20th century. This development is summarised in Figure 2. We identify three steps in this evolution: starting with H2H (Human-to-human), to H2S (Human-to-System), and finally to S2S (System-to-System). There is, however, no clear distinction between these steps in time; in fact nowadays all methods are applied. The second and third steps have their origins in the 1980s, while the first step goes back to the early 1900’s when survey sampling became the common approach, and even to the 1800’s when in many countries National Statistical Institutes were established (using a census approach). For a more detailed discussion on the development in digitalisation in business data collection since the 1980’s we refer to Buiten et al. (2018).

**Figure 2. Evolution of digitalisation in primary business data collection** (adapted from Buiten et al., 2018: figure 1)

In addition to primary data collection, in the 1970s secondary data collection emerged: registers and administrative data were being used in the production of statistics, to replace surveys, supplement survey data, or for data validation or imputation. Because of the computerisation of registers, the exchange of large data files became feasible, and NSIs had to develop multi-source approaches to combine primary and secondary data. This process is still going on. Another secondary data source that nowadays gets a lot of attention is big data (Groves 2013). Since 2012 the potentials of big data in official statistics are being explored (see e.g. Di Consiglio 2016; Daas et al., 2015). In section 3 we will discuss new business data sources.

These developments are shown in figure 3. With all these new data sources, in follow-up to Zhang (2012), we could state that the 21st century will be age of data integration, while the 20th century is characterised as the age of sampling. The consequence for an NSI is that they will be more and more out of control of the actual data collection. If NSIs do not pay attention, they will be followers with regard to the data that serve as input for their production processes.

**Figure 3. From single source to multiple sources**



**3. Enablers: Innovations in the business world**

In the business world we see innovative developments that NSIs need to consider seriously: businesses are moving toward Intelligent Business Architecture. This includes further automatisation and application of smart sensing technology, the explosive growth in information, and on-going real-time integrated data communication via the Internet of Things (Srinivasan, 2017; Hinssen, 2017; Vermeend & Timmer, 2016; Bharosa et al., 2015), at two levels:

1. Intelligent robotisation and integration of business production processes
2. Digitalisation and integration of administrative business processes

Below, we will illustrate these two levels with examples.

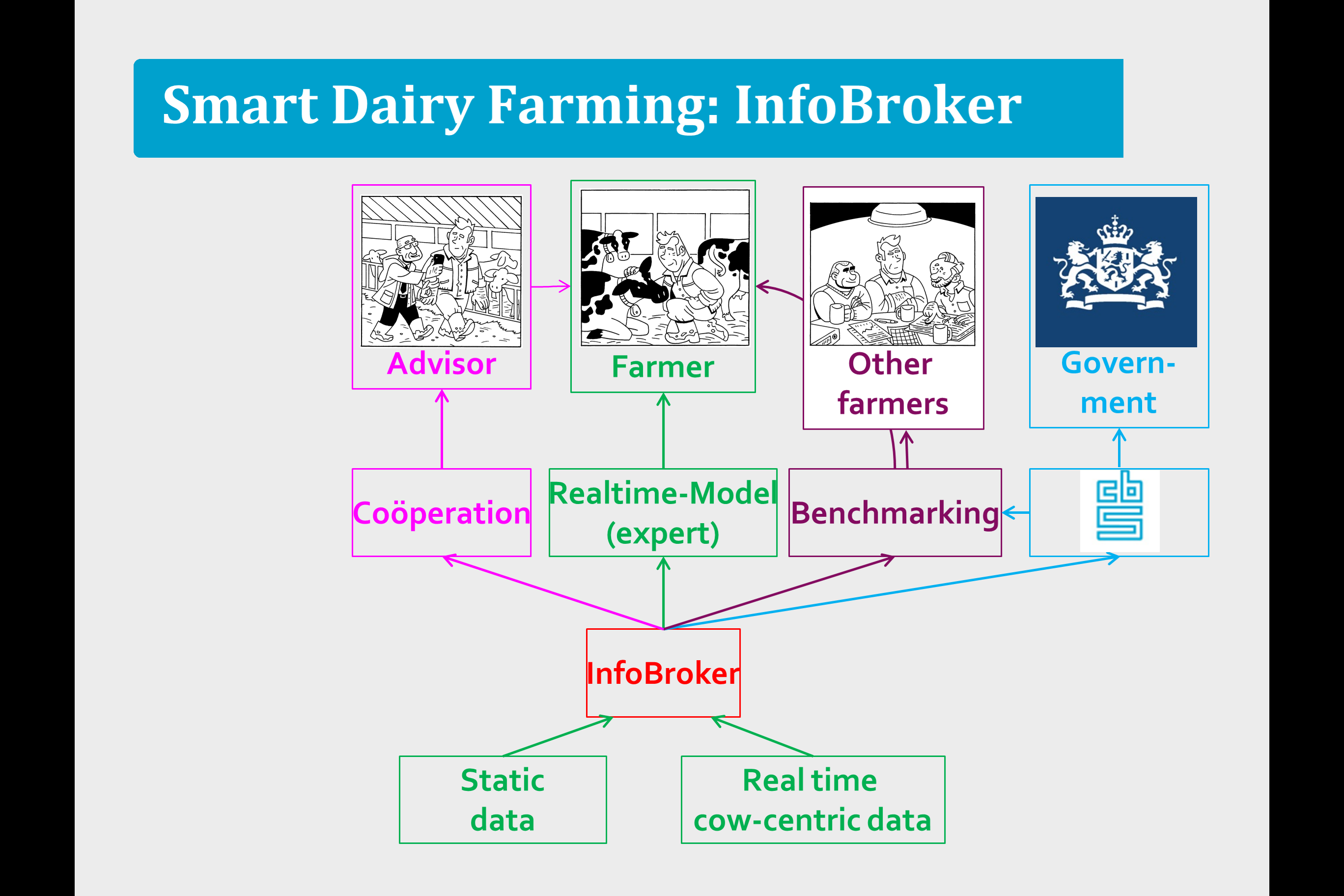
As to the first development an example is smart and precision farming. The agricultural sector is innovating quickly in the direction of smart sensing and monitoring (De Vlieg, 2018; EU-AOC, 2017; Wolfert et al., 2017; Thomas & McSharry, 2015). Examples include:

* Dairy farmers using sensor data to run their business. Dairy farmers use milking robots with sensors, generating sensor data like milk production per cow. When cows are being milked, cow-related variables are being measured, like weight, temperature, and quality of the milk. These data feed into an app on the farmer’s cell phone, which enables him to monitor the milk production and his life stock in real time (Vonder, 2017; Brandt Corstius, 2017).
* Potato farmers using intelligent technology to monitor their land and crop. Precision farming here includes satellite data and sensor data collected by drones and smart machines (like driverless tractors) to monitor humidity and fertility of the soil, showing where additional irrigation and fertilisation is needed (Viviano, 2017).
* Greenhouse horticulture is another agricultural sector in which smart sensing and monitoring is used to run the business, e.g. the temperature and humidity in greenhouses, the energy consumption, and the growth of every plant.

An overview of open data for precision farming in the Netherlands, as provided by van Dijk and Kempenaar (2016), includes satellite, meteorological, soil, crop data.

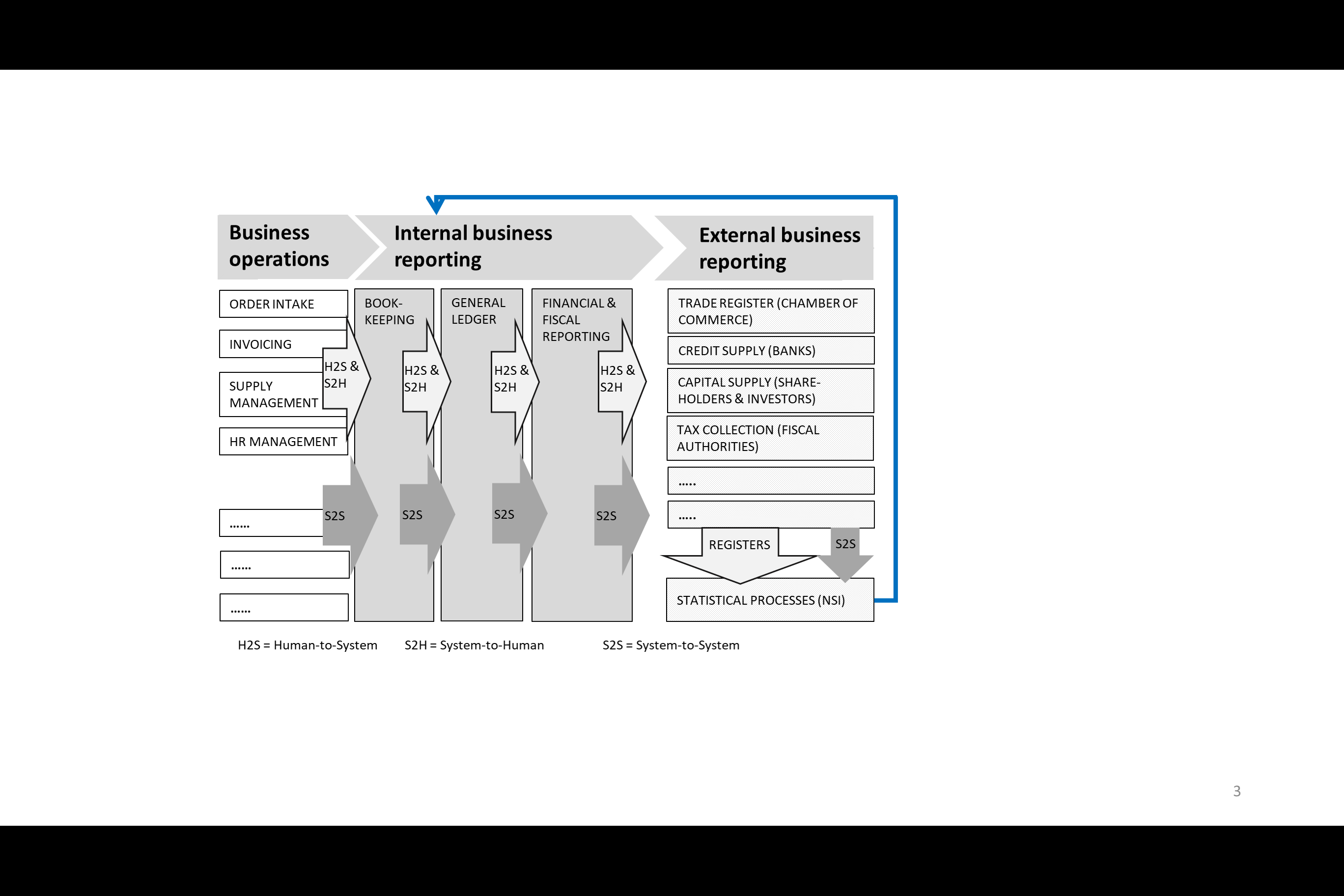
Despite the technological infrastructure being available, and farmers having invested in sensing technology, farmers may not have the capabilities to optimally use these data (de Vlieg, 2018). To assist farmers in data analysis and usage, data service centers (DSC) are being established (van Dijk and Kempenaar, 2016). Farmers can become a member of such a DSC. These DSCs serve as data hubs, and assist their members in monitoring their business: they collect all data from their members, integrate these data in one overall data base, also including open data from other sources, analyse the data, and develop apps for business monitoring and dashboarding. Apart from monitoring individual businesses, the data also can be used for benchmarking (over all members). These functions of a DSC are shown in figure 4, making an integrated unbroken digital data chain possible (figure 2).

In the Netherlands a DSC model is developed by the Dutch Organisation for Applied Research TNO in the Smart Dairy Farming project ([www.smartdairyfarming.nl](http://www.smartdairyfarming.nl); Vonder, 2017; Van Dijk and Kempenaar, 2016): InfoBroker for dairy farmers. The InfoBroker model and its uses is shown in figure 4. In 2017 this InfoBroker has been extended to a company called JoinData ([www.join-data.nl](http://www.join-data.nl)), which has the ambition to become a DSC for the entire agricultural sector.

**Figure 4: Smart Dairy Farming: Some scenario’s for using the InfoBroker**(adapted from Vonder, 2017)

In addition to the agricultural sector, also other industries are innovating in the direction of intelligent robotisation and usage of sensing and monitoring technology like smart manufacturing industries and the transportation sector.

Our second example relates to administrative business processes: digitalisation and integration in financial business reporting, both internally and externally, working toward a fully integrated and unbroken business information chain (Bharosa et al., 2013; Buiten et al., 2018). The business information chain, as represented in figure 5, shows the internal and external exchanges of administrative data at various aggregation levels. The chain starts with business operations generating data at transaction levels such as product codes, shipment numbers, bar codes, customer IDs, etc. These data may be stored in one or more subsidiary ledgers such as customer administration, stock administration, wage administration, and so on, which can reside in different business departments, and are input for administrative processes through which the business translates data into dashboards, financial reports, tax declarations, etc. (Bharosa 2015; Wolfert et al., 2013; Zhang 2012). For this business information chain, businesses are moving from H2S/S2H data communication toward integrated S2S data communication. (See Buiten et al., 2018.)

**Figure 5. The integrated business information chain** (from Buiten et al., 2018)

In the context of developing integrated business information chains, we also see the establishment of DSCs. E.g. Dutch business sector organisations are establishing DSCs for their SME members, like for restaurants and pubs. These DSCs serve the same purposes as discussed above for the agricultural sector.

**4. Challenges: Consequences for collecting business data**

The developments discussed above open up a whole new world of data. They also require NSIs to rethink traditional business data collection (as discussed by Snijkers et al., 2013). The question we need ask ourselves is: why do we still use questionnaires while the data are already electronically available. In other words: why not innovate our business data collection using EDI technologies? This question has not been posed for the first time (Buiten et al., 2018), but now is the time to discuss this again.

Effective application of EDI technology, making S2S data communication possible to collect large amounts of business data like sensor data and financial data, however requires some methodological and organisational issues to be solved. In the introduction we stated that technology is the enabler of innovations; it is the applied methodology and organizational context that make innovations work. We will now discuss the challenges we face when implementing EDI. The major issues include:

* Harmonisation of data and metadata. When integrating data from several sources in one data base in a DSC, data definitions need to be harmonised. E.g. sensor data generated by milking robots from various producers, need to be defined in the same way.
* IT infrastructure. Assuring the interconnectivity and interoperability between all stakeholders through all steps in an integrated business data chain, technical standards have to be put in place that are easily and widely accepted.
* Conceptualisation. When using existing data for the production of regulated statistics, as opposed to designed data (Groves, 2013), these data need to be linked to concepts for which statistics are to be produced. In surveys these concepts are operationalised in questionnaires (Snijkers et al., 2013), resulting in designed data. In the case of existing data we need to rethink how the existing data correlate with the concepts to be measured.
* Unit issues. When collecting data automatically from businesses, a major concern is the unit that is associated with these data. Especially for larger businesses, the statistical unit may be composed out of a number of smaller units. The reporting business unit may need to collect data from these smaller units and add the data together, to get the required consolidated data.
* The use of sampling. With the application of EDI, NSIs need to rethink sampling. For businesses it is easier and cheaper to invest in a system and then deliver the data by literally pressing a button, then delivering the data on the basis of a sample. This would mean that we move to integral data collection. as was the case before the introduction of sampling.
* Data quality issues and data cleaning. Data quality is a very broad concept and includes e.g. measurement errors, non-response, representativeness, but also the above mentioned unit issue (Snijkers, 2016). In terms of sensor data quality one can think of repeatability (the dispersion between consecutive measurement obtained from a given sensor), reproducibility of the measurements with different sensors, stability of the sensor and data drift (the capability of the sensor to maintain its performance characteristics over a sufficiently long time) and limit of detection (the lowest concentration of a substance that can be significantly differentiated from zero concentration) (Rai et al., 2017). To check for errors in the data, data cleaning is an important part of the data process.
* Data-ownership, privacy and security. These issues are of another level as the ones above, but of major importance. When lots data are shared, these legal issues need to be considered carefully. This has to do with trust in the data users, and resistance by the data owners in sharing the data.
* Stability of (meta)data delivery. When the production of official statistics becomes dependent on existing data, the delivery of these data should be guaranteed, as well as the structure of the data and the metadata.
* Costs, data storage, maintenance. Both businesses as well as an NSI need to invest in these intelligent systems, while the revenues of these investments may not be clear upfront (De Vlieg, 2018). Issues to be dealt with include data storage and maintenance of the system.
* Resistance. Within an NSI, since the data are produced out of our control, resistance against the use of these new data may be expected.

These challenges relate to the FAIR data principles: Findable, Accessable, Interoperable, and Reusable (Wilkinson et al., 2016). Many of them are related to the data generating process. Some are not new; they have also been discussed in the context of using register data (Daas et al., 2011).

**5. Moving toward smart statistics**

Statistics Netherlands (SN) has defined an innovation strategy with a clear focus on making policy relevant statistics based on new methods and new secondary data sources. Sensor data are one of the data sources that SN will aim to fully exploit for the use of official statistics. Statistics Netherlands has already explored the use of sensor data as in traffic loop data for the traffic intensity statistic, aerial pictures using machine learning for detecting solar panels and satellite data for air pollution data.

Sensor data offer great potential but for an NSI the challenge lies in defining the use from the end-user perspective. In order for sensor data to be of any use for end users (and to become part of an official statistical process) many issues as discussed in section 4 still have to be dealt with.

As a starting point for investigating the use of business sensor data in business statistics the Methodology Department of Statistics Netherlands opted for the agricultural sector as a first candidate for a pilot. The arguments for this decision were twofold. First of all getting high enough response rates for agricultural surveys has always been cumbersome. Secondly the above discussed developments in the agricultural sector seem promising from an official statistics perspective.

The main drivers for this research are fourfold:

* to use these new data sources to produce statistics in order to reduce response burden and potentially costs,
* to develop a new methodology, including new techniques and methods dealing with these new data sources and the above mentioned challenges,
* to produce more detailed, more timely and potentially new statistics, and
* to close the business data chain and turn it into a data cycle. Instead of having the data flow go in one direction (from businesses to DSC to NSIs), we can provide relevant statistical information back (for benchmarking). Instead of just collecting data, businesses will finally get something in return, thus addressing their question: “what is in it for me?”

In a fully integrated data cycle near-real time statistics could be produced. The benchmark information could be produced by DSCs themselves by accessing open data. This would be ‘smart statistics’ (Eurostat, 2017, p. 4): “… the future system of official statistics, where data capturing, processing and analysis will be embedded in the system itself, starting with the digital footprints of activities”.

To conclude this paper, in Section 4 we saw that many issues still need to be addressed in order to make EDI and the use of already electronically available data work. In section 2 we concluded that with the increased use of secondary data NSIs lose control over the data collection. In order to gain some control and address the challenges, we foresee another business model for NSIs like Statistics Netherlands: be a stakeholder and partner in emerging DSCs (as discussed in Section 3) from the very start, and be in close communication with the end-user (e.g. the farmer in our pilots) so that data and output are of utmost relevance. The role of an NSI in these DSCs should not only be to collect data, but also enrich these DSC data with relevant statistical information, and assist in the use of statistical information. Now is the time to establish these contacts, not missing the DSC boats! It is our view that:

* Sensor data and fully integrated business data chains offer great potential for NSIs; however, in the first instance methodological issues, including data quality issues, will have to be dealt with.
* Legal routes have to be defined in order to protect privately sensitive data whilst at the same time allow NSI access to private data in order to reduce response burden and make policy relevant statistics.
* Early intervention as a stakeholder is of utmost importance for determining data collection processes as well as relationship with the end user.

So, despite the fact that technology is the enabler of innovations, it is the applied methodology and organisational challenges that make innovations work.

**5. References**

Brandt Corstius, J. (2017), Robo Sapiens: Work. TV documentary series, part 3. VPRO Broadcasting Company, Hilversum, Netherlands.

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*, Vol. 34, No. 2 (ICES-5 special issue).

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren (2011), List of quality groups and indicators identified for administrative data sources. Luxembourg: Eurostat (BLUE-ETS deliverable 4.1).

Daas, P., M. Puts, B. Buelens, and P. van den Hurk (2015), Big Data as a Source for Official Statistics. *Journal of Official Statistics* 31(2): 249-262.

De Vlieg, J. (2018), ‘A huge push in technology is coming’ (in Dutch: ‘Er is een enorme push in technologie in aantocht’). Boerderij, 03: 19 (februari 2018).

Di Consiglio, L., M. Karlberg, M. Skaliotis, I. Xirouchakis (2016), Overview of Big-Data research in European Statistics Agencies. In Proceedings of the Fifth International Conference of Establishment Surveys: Statistics (ICES5) Switzerland, 20-23 June 2016, Geneva, Switzerland. Statistics Switzerland.

Eurostat (2017), Smart Statistics. Paper discussed at the Joint Dime/ITDG plenary sessions, 14-15 February 2017. Luxembourg: Eurostat.

Groves, R.M. (2013), Official Statistics and “Big Data”. Paper presented at the 2013 European Conference for New Techniques and Technologies for Statistics (NTTS), Brussels, 5-7 March 2013. Luxembourg: Eurostat.

EU-AOC (2017), The 2017 EU Agricultural Outlook Conference. Brussels, 18-19 December 2017 (<https://ec.europa.eu/info/events/2017-eu-agricultural-outlook-conference_en>).

Hinssen, P. (2017), *The day after tomorrow: How to survive in times of radical innovation.* Lannoo, Tielt.

Rai, A., P. Kumar, F. Pilla, A. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, and D. Rickerby (2017), End-user Perspective of Low-cost Sensors for Outdoor Air Pollution Monitoring. *Science of The Total Environment*, 607-608: 691-705.

Snijkers, G. (2016), Achieving Quality in Organizational Surveys: An Holistic Approach. In *Methodische Probleme in der empirischen Organisationsforschung*, S. Liebig, and W. Matiaske (eds.), 33-59. Wiesbaden: Springer.

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

Srinivasan, V. (2017), *The Intelligent Enterprise in the Era of Big Data*. Hoboken, NJ: Wiley.

Thomas, R., and P. McSharry (2015), *Big Data Revolution: What Farmers, Doctors, and Insurance Agents teach us about discovering Bog Data Patterns*. Wiley, Chichester, West Sussex, UK.

Van Dijk, C., and C. Kempenaar (2016), Open data for precision farming in the Netherlands (in Dutch: Open data voor precisielandbouw in Nederland). Wageningen Univeristy & Research, report 662. Wageningen, Netherlands.

Vermeend, W., and J.W. Timmer (2016), *Internet of Things: Technology that will change the world* (in Dutch). Einstein books, The Hague.

Viviano, F. (2017), This tiny country feeds the world. *National Geographic*, September 2017. ([www.nationalgeographic.com/magazine/2017/09/holland-agriculture-sustainable-farming/](http://www.nationalgeographic.com/magazine/2017/09/holland-agriculture-sustainable-farming/)).

Vonder, M. (2017), Sensors going smart. Presentation at ‘Big Data Matters’ Seminar, Statistics Netherlands, 27 September 2017, Heerlen, Netherlands. (TNO Netherlands)

Wilkinson, M.D., et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 15 March 2016.

Wolfert, S., L. Ge, C. Verbouw, M.-J. Bogaardt (2017), Big Data in Smart Farming – A review. *Agricultural Systems* 153: 69-80.

Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* 66(1): 41–63.

1. Disclaimer: The views expresses in this paper are those of the authors and do not necessarily reflect the official policy of Statistics Netherlands. [↑](#footnote-ref-1)