CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection 'New Sources and New Technologies' Session 3

14-16 October 2019, Geneva, Switzerland 20 September 2019

Future Advanced Data Collection

Irene Salemink (Statistics Netherlands), Stéphane Dufour (Statistics Canada) and Marcel van der Steen (Statistics Netherlands

islk@cbs.nl

Abstract and Paper

Society's demand for data-driven, fact-based information continues to increase. National statistical offices (NSOs) play a critical role in providing this demand-driven information to support evidence-based policy making. In a datafied society, NSOs are transforming from suppliers of official statistics to providers of trusted smart statistics, coping with various aspects of complexity and keeping a keen eye on fast-emerging trends.

When it comes to data, collecting data and producing statistics, the times have dramatically changed for NSOs. The digital transformation, data revolution and emergence of "big data" all influence the way NSOs collect data. Data are everywhere, generated by everything and everyone. These data are stored in numerous locations and devices. The enormous increase in computing power and cost-efficient storage capacity have created never-before-seen analytical performance capabilities.

This development sets new requirements for data collection. Sticking to conventional methods based on sampling theory and using primary data collection would be too time-consuming, costly and burdensome to satisfy the increasing demand. NSOs should aim to use the vast amounts of data that are or that become available in our digital society. These data can be used as inputs for new statistical products, that respond best to users information needs to supplement existing data acquisition or as replacements for existing survey inputs in the statistical production process.

Consequently, the nature of data collection is bound to change. Using administrative data and sensor data is a logical step toward the future of data collection. Many areas must be taken into account to bring out the full future potential of NSOs, including new data sources, new collection methods and collection process redesigns. All these options come with certain consequences with respect to methodology, technology, quality, metadata and other standards, confidentiality, privacy, acceptability, and more. The amount of knowledge development required calls for extensive collaboration not only between NSOs, but also with governments, end users, academic institutions, research organizations and private sector companies. Social acceptability also needs to increase to maximize the benefit of these data sources to produce smart statistics.

To be able to do that, it is necessary to develop a new way of looking at data collection and at the official statistical production process as a whole. Instead of acting from specific variables linked to a dedicated collected dataset, NSOs need a paradigm shift. The focus needs to be on reviewing available data, determining what is missing (gaps), and searching for opportunities to design new outputs or to replace part of the existing output.

As a consequence, surveys are likely to become smaller and more tailored, and to be used increasingly for validation purposes. A new observation strategy has emerged as a result of new data sources and observation

techniques becoming available. This strategy includes removing as much of the burden of primary data collection from society as possible. The vision paper on Future Advanced Data Collection describes strategic assignments for NSOs, relevant trends that are changing the nature of data collection, challenges and constraints regarding data capture, and an envisioned outline of future data collection.

Finally, a call to action is made as joint efforts in open innovation and further co-creation partnerships and collaboration between the statistical sector, universities, the private sector and end users will be necessary in the following areas:

- Methodology: Develop new statistical methods or adapt existing ones to deal efficiently with using administrative data and sensor data in areas such as data linkage, data integration and data validation.
- Collect, connect and link: Accelerate the development of techniques and technology to facilitate access to new data sources and ensure even greater confidentiality and privacy protection (PPDS, PPRL, multi-party computation, edge computing, etc.).
- Metadata: Develop and promote structures and standards to facilitate the evaluation, curation and linkage of survey data, administrative data and proprietary sensor data.
- Proprietary sensor networks: Sensor networks dedicated to statistical purposes need to be designed, developed and tested so that standards and guidelines can be developed and efficient deployment can be accelerated.
- Legal frameworks and social acceptability: Legal frameworks need to be developed (or existing ones deepened) to establish responsible and ethical access to administrative and sensor data in such a way that confidentiality and privacy are protected and social acceptability is obtained.

Furthermore, showcases need to be developed to demonstrate the possibilities and sustainability of the new methods and techniques in the areas listed above.

Open innovation and partnerships

Small targeted coalitions and strategic partnerships are needed to work on these items to unleash the added value of future advanced data collection. CBS and Statistics Canada have taken the initiative for this vision paper and are prepared to coordinate achieving this goal with all interested organizations that are able and willing to contribute.

Keywords

[Keywords]

Statistics Canada Statistique Canada

Vision Paper on

Future Advanced Data Collection

Irene SALEMINK
Director, Innovation, Development and Support
Statistics Netherlands (CBS), Heerlen, Netherlands
Stéphane DUFOUR
Assistant Chief Statistician, Census, Regional Services and Operations Field
Statistics Canada, Ottawa, Canada
Marcel van der STEEN
Chief Innovation and Strategic Partnerships Officer
Statistics Netherlands (CBS), The Hague, Netherlands

August 2019

EXECUTIVE SUMMARY

Society's demand for data-driven, fact-based information continues to increase. National statistical offices (NSOs) play a critical role in providing this demand-driven information to support evidence-based policy making. In a datafied society, NSOs are transforming from suppliers of official statistics to providers of trusted smart statistics, coping with various aspects of complexity and keeping a keen eye on fast-emerging trends.

When it comes to data, collecting data and producing statistics, the times have dramatically changed for NSOs. The digital transformation, data revolution and emergence of "big data" all influence the way NSOs collect data. Data are everywhere, generated by everything and everyone. These data are stored in numerous locations and devices. The enormous increase in computing power and cost-efficient storage capacity have created never-before-seen analytical performance capabilities.

This development sets new requirements for data collection. Sticking to conventional methods based on sampling theory and using primary data collection would be too time-consuming, costly and burdensome to satisfy the increasing demand. NSOs should aim to use the vast amounts of data that are or that become available in our digital society. These data can be used as inputs for new statistical products, that respond best to users information needs to supplement existing data acquisition or as replacements for existing survey inputs in the statistical production process.

Consequently, the nature of data collection is bound to change. Using administrative data and sensor data is a logical step toward the future of data collection. Many areas must be taken into account to bring out the full future potential of NSOs, including new data sources, new collection methods and collection process redesigns. All these options come with certain consequences with respect to methodology, technology, quality, metadata and other standards, confidentiality, privacy, acceptability, and more. The amount of knowledge development required calls for extensive collaboration not only between NSOs, but also with governments, end users, academic institutions, research organizations and private sector companies. Social acceptability also needs to increase to maximize the benefit of these data sources to produce smart statistics.

To be able to do that, it is necessary to develop a new way of looking at data collection and at the official statistical production process as a whole. Instead of acting from specific variables linked to a dedicated collected dataset, NSOs need a paradigm shift. The focus needs to be on reviewing available data, determining what is missing (gaps), and searching for opportunities to design new outputs or to replace part of the existing output.

As a consequence, surveys are likely to become smaller and more tailored, and to be used increasingly for validation purposes. A new observation strategy has emerged as a result of new data sources and observation techniques becoming available. This strategy includes removing as much of the burden of primary data collection from society as possible. The vision paper on Future Advanced Data Collection describes strategic assignments for NSOs, relevant trends that are changing the nature of data collection, challenges and constraints regarding data capture, and an envisioned outline of future data collection.

Finally, a call to action is made as joint efforts in open innovation and further co-creation partnerships and collaboration between the statistical sector, universities, the private sector and end users will be necessary in the following areas:

- **Methodology:** Develop new statistical methods or adapt existing ones to deal efficiently with using administrative data and sensor data in areas such as data linkage, data integration and data validation.
- **Collect, connect and link:** Accelerate the development of techniques and technology to facilitate access to new data sources and ensure even greater confidentiality and privacy protection (PPDS, PPRL, multi-party computation, edge computing, etc.).
- **Metadata:** Develop and promote structures and standards to facilitate the evaluation, curation and linkage of survey data, administrative data and proprietary sensor data.
- **Proprietary sensor networks:** Sensor networks dedicated to statistical purposes need to be designed, developed and tested so that standards and guidelines can be developed and efficient deployment can be accelerated.
- **Legal frameworks and social acceptability:** Legal frameworks need to be developed (or existing ones deepened) to establish responsible and ethical access to administrative and sensor data in such a way that confidentiality and privacy are protected and social acceptability is obtained.

Furthermore, showcases need to be developed to demonstrate the possibilities and sustainability of the new methods and techniques in the areas listed above.

Open innovation and partnerships

Small targeted coalitions and strategic partnerships are needed to work on these items to unleash the added value of future advanced data collection. CBS and Statistics Canada have taken the initiative for this vision paper and are prepared to coordinate achieving this goal with all interested organizations that are able and willing to contribute.

Statistics Canada Statistique Canada

Future Advanced Data Collection

Irene SALEMINK, *Director, Innovation, Development and Support, Statistics Netherlands (CBS), Heerlen, Netherlands*
Stéphane DUFOUR, *Assistant Chief Statistician, Census, Regional Services and Operations Field, Statistics Canada, Ottawa, Canada*
Marcel van der STEEN, *Chief Innovation and Strategic Partnership Officer, Statistics Netherlands (CBS), The Hague, Netherlands*

Society's demand for data-driven, fact-based information continues to increase. National statistical offices (NSOs) play a critical role in providing this demand-driven information to support evidence-based policy making. In a datafied society, NSOs are transforming from suppliers of official statistics to providers of trusted smart statistics, coping with various aspects of complexity and keeping a keen eye on fast-emerging trends.

This development sets new requirements for data collection. Sticking to conventional methods based on sampling theory and using primary data collection would be too time-consuming, costly and burdensome to satisfy the increasing demand and potentially underuse the great potential of new secondary data sources containing new information. Consequently, the nature of data collection is bound to change. Individuals, organizations and non-living objects generate a multitude of data, and technology and techniques to retrieve and process data are continuously evolving. Therefore, using administrative data and sensor data is a logical step toward the future of data collection. Innovative methods, legal frameworks and technology need to be further developed. Social acceptability also needs to increase to maximize the benefit of these data sources to produce smart statistics.

Guaranteeing confidentiality and privacy will require intensive cooperation between NSOs, partnerships with policy makers, and strategic partnerships with data and technology providers and researchers. Together, we can pave the future pathway of advanced hybrid data collection. This paper describes strategic assignments for NSOs (§1), relevant trends that are changing the nature of data collection (§2), challenges and constraints regarding data capture (§3), and an envisioned outline of future data collection (§4). The paper concludes with a call to action for the necessary collaboration (§5).

1. Strategic assignments for national statistical offices

Currently, the mission of NSOs is to provide end users with trusted official statistics that they can use to understand, monitor and manage the economy and society. The majority of these official statistics are based on a mandatory program consisting of a set of consensus indicators that describe economic, social and demographic phenomena. However, this mission is bound to change to create more added value to society.

1.1. Demand driven

The value of an NSO lies in its ability to reliably deliver a broad range of fit-for-purpose statistics that correspond to user needs, with an acceptable degree of accuracy and in a sufficiently timely fashion.[1] This

1. Bean, C. 2016. *Independent Review of UK Economic Statistics.*

information is vital not only for policy makers, but also for effective decision making in the private sector and opinion forming in society. Moreover, for media and the public, access to reliable, relevant, accurate and timely statistics is necessary for holding decision makers accountable. However, it is becoming evident that these phenomena are often so complex that using a single indicator or a limited set of indicators does not provide enough information or accurate information for this purpose. No single set of statistics is likely to cover all purposes; different users have different needs. Delivering these sets of consensus indicators to users as periodic publications that are released at fixed times also does not meet the criterion of "information upon request, when needed."

To become more valuable to society, NSOs must be able to answer society's statistical questions without compromising their independence, transparency and freedom to determine the scientific methods used. The conclusion is clear: **being user centric and demand driven must be a critical transformation pillar for NSOs**.

1.2. Evidence-based policy making

Policy makers are becoming more and more aware that the availability of an increasing amount of data offers huge possibilities for evidence-based policy making. At the same time, it is clear that the current set of consensus indicators that NSOs are producing within their mandatory statistical programs generally does not suffice. Policy makers need access to more timely data products, indicators and insights, and they are looking for **actionable intelligence**, i.e. high-content information that can be acted upon immediately. **Official statistics are a means to that end.**

This may raise questions about a possible key collective goal for NSOs: to develop and provide access to more self-serve tools with ready-to-consume "cleansed" data that NSOs have collected, curated, protected and certified, then make the data available at various levels of aggregation. Does this mean that NSOs must become more adept in data preparation? Are NSOs currently unable to be demand driven and user centric, and unable to support evidence-based policy making, because they take too long to conduct analyses and share analytical results and insights? The following paragraphs address several aspects of complexity related to the use of official statistics for evidence-based policy making and decision making.

1.3. Dealing with various aspects of complexity

1.3.1. Increasingly complex societal and economic phenomena

Adequately measuring societal and economic phenomena in an ever-changing world is not self-evident, especially not in this era of increasing digitalization, globalization and shift toward services. Leading economists have determined that the observed decrease in productivity growth is probably partially caused by the inability to measure accurately the increased influence of the digital economy. Online shopping, digital platforms and a wide range of digital services delivered through apps are good examples of focus areas for NSOs.

In addition, there is general agreement that the world economy is becoming more integrated and interdependent economically, culturally and politically[2]. This makes it challenging to describe phenomena from a national perspective.

NSOs will have to invest in statistical resources to fully characterize and better respond to economic globalization, digitalization and the shift toward services. Furthermore, it is key that **NSOs join forces with other international partners** to develop expertise in these areas and to create access to new data sources that will allow for socioeconomic phenomena to remain measurable and indicators to remain relevant.

2. For example, multinationals play an increasingly large role in n the international economy. With their sister companies or subsidiaries it is hard to see how the associated shifts of capital, intellectual property and internal transfer pricing affect the figures of a national economy.

1.3.2. Real-time statistics

Statistical information end users expect information that addresses their specific needs. Timeliness of information is a key aspect of their needs. End users expect to be informed as soon as possible of events occurring in society and their potential consequences.
Currently, NSOs continuously provide information on society and cover a wide range of topics. Although this information is highly accurate, the timeliness—and, therefore, usefulness—of the information is limited because of the nature of the raw material (data) and the statistical production methods.

Also, not every method is yet sufficiently robust to apply event-driven processing (EDP). The method of raising respondents to a population requires enough responses for it to be applied. To apply EDP-like approaches instead, a possible solution could be to initially impute non-responses with historical data, and then replace the historical data with actual data once they are received. Various statistical methods will need to be adapted to introduce EDP and become more responsive to actual events. Furthermore, the **success** of this kind of **EDP initiative** remains highly **dependent** on the **timely acquisition of administrative and other alternative data**.

Examples:
One example is a method that was developed to estimate economic demography using the daily registration and deregistration of businesses with the chamber of commerce (or other responsible institutions) and updating the statistical business register. The method proves that it is possible for a statistical process to react to real daily events in society and estimate, on a daily basis, the society's enterprise population.

Other examples of real-time information on societal events include the Dutch business cycle tracer[3], the population counter[4] and the AG-Zero initiative being explored at Statistics Canada. The goal of the AG-Zero initiative is to provide objective, high-quality statistical information for the agriculture industry while reducing response burden and ensuring minimum delays to information releases and minimum data suppression. This approach is based on eliminating (or keeping to a strict minimum) direct contact with respondents by using alternative sources of information. Building solid partnerships with data providers and users is key. By adding value for users through the additional information and products provided, and by offering statistical capacity building and sharing expertise, these partnerships should facilitate the timely sharing of required data.

Accuracy and timeliness are often seen as conflicting priorities—increased accuracy of statistical information leads to decreased timeliness. Although this might be true to some extent, it is important to consider these two parameters and discuss them together. On the one hand, future advanced data collection will need to reduce this conflict. On the other hand, fit-for-purpose statistics will need to be discussed. As long as the accuracy level of the information provided is known, timeliness can be more important than accuracy from an end user's perspective.

1.3.3. Tailored to all aggregation levels

The data collected and processed by NSOs are used for various purposes and by various end users. Governments and enterprises often base their policy making and decision making on NSO information. Universities use statistical microdata for research, and citizens can benefit from NSO statistics on their neighbourhoods.

In many countries, if not all, local governments are facing an intensified need for local policy making. They also realize that using data can support evidence-based policy making. Therefore, local authorities need reliable regional statistical information.

3. https://www.cbs.nl/en-gb/visualiaties/business-cycle-tracer
4. https://www.cbs.nl/en-gb/visualisaties/population-counter

For example, the need for neighbourhood statistics on the percentage of households with long-term low income, the workforce participation rate, educational levels of the labour force, or economic activity is almost self-evident. Itemizing national statistical figures into various aggregation levels (e.g., international, national, regional, local) is necessary, but insufficient. There is a **need for more detailed and custom-fit regional and local information** on a variety of topics such as daytime population, mobility (e.g., movements, time, numbers, and mode of transportation), energy use per economic sector (e.g., service industry, industry, households) and disused premises.

The Sustainable Development Goals (SDGs) are another example of this increased complexity.[5] Naturally, the "sustainable cities and communities" target will directly affect local policy making. However, the SDGs have far-reaching consequences for public sector institutions. SDGs play a key role in designing national development strategies and plans; monitoring and evaluating public programs, projects and development activities; giving public servants the tools to implement the global SDGs; and providing inclusive services to better serve local citizens through digital government.

Like cities and municipalities, NSOs also face challenges to becoming data driven. **It is impossible for NSOs to produce highly detailed, complex information solely through the traditional survey-based data collection approach.** It is too time-consuming, expensive and burdensome, and it is increasingly exacerbated by a decrease in survey participation.[6]

Future advanced data collection will entail using other data sources (e.g., administrative data, sensor data) to fulfill the need for tailored information for all aggregation levels of society. Again, NSOs will have to move away from standard, one-size-fits-all quality and toward **"fit-for-purpose" quality**. Quality has several dimensions, and the data's intended use should determine the criteria for accuracy, relevance, timeliness, bias, sample and processing cost.

1.4. Shaping the future by being smart

NSOs have a leading role in shaping the future of official statistics. In a datafied society and economy, official statistics are evolving into smart statistics through the introduction of smart technologies and smart data. By guaranteeing confidentiality and privacy by design, NSOs should aim to become providers of **trusted smart statistics**. This will deliver facts that can be used for evidence-based policy making and for quantitative monitoring of developments and progress; that are society oriented, reliable and innovative; and that protect confidentiality and privacy.

The growing end-user demand for accurate and up-to-date smart statistics forces NSOs to constantly evolve to be able to produce statistics with more efficiency and less administrative burden. It is clear that **this goal cannot be achieved by relying solely on conventional methods and the traditional survey-based approach**. Around the world, traditional primary data collection methods are becoming less effective and require more effort to achieve satisfactory results.

Technological and cultural changes have increased collection costs because establishing contact with respondents and gaining their cooperation now require more effort. As a result, response rates for many surveys are trending downward. Finding new, innovative ways to collect the data required for smart statistics is key. **Consequently, the nature of primary data collection for trusted smart statistics is bound to change.** The quantity of detailed, real-time data generated by sources such as social media, road sensors, smartphones and the Internet is always increasing. These data are more flexible and detailed and, therefore, create a huge opportunity to enhance the value of official statistics. With advanced information technology (IT), these data can be analyzed more rapidly and in greater depth. By employing innovative methods, NSOs can tap into these new data sources and deliver the right information to the right people at the right time and in the right format. Using multiple, interrelated indicators to clarify and indicate complex social and economic phenomena creates

5. This year's theme on the United Nations Public Service Forum; see https://publicadministration.un.org/en/UNPSA2019.
6. Jarmin, R. 2019. "Evolving measures for an evolving economy: Thoughts on 21st century US economic statistics." *Journal of Economic Perspectives* 33 (1), 165–184.

unprecedented possibilities for governments and for other social actors to develop better, more targeted and more effective fact-based policies. These include, for example, new policy measures and legislation, smart city concepts, and individual decisions by businesses and private citizens.

1.5. New business models

It is clear that the demand for timely, factual information about complex societal phenomena is rising. This demand exceeds what the current national mandatory statistical programs can provide in terms of information content, aggregation level and timeliness. These product requirements will give rise to new business and funding models for the statistical sector, with respect to the timely acquisition of the necessary raw data needed to produce the requested information. To fulfill the demands of faster and higher-resolution statistics (at lower aggregation levels), using primary data alone will not suffice. Therefore, the development of new business models to provide administrative and sensor data is necessary.

2. Trends

When it comes to data, collecting data and producing statistics, the times have dramatically changed for NSOs. The digital transformation, data revolution and emergence of "big data" all influence the way NSOs collect data. Data are everywhere, generated by everything and everyone. These data are stored in numerous locations and devices. The enormous increase in computing power has created never-before-seen analytical performance capabilities. The datafication of society, the availability of cheap data storage and the availability of analytical power are the main trends driving the next steps in sensor data collection.

2.1. Datafication

Data are a vital part of daily life. Planning a route, uploading content to social media, creating personalized sales items, taking public transit, entering an office building with an access card—all these actions create data and leave a digital trail. Data are always influencing society and are essential for a functioning society, an effective and transparent government, innovation, and economic activity. This datafication of society also makes many new data sources available.[7]

Although administrative data[8] and sensor data make producing statistics potentially more complex, conceptually and methodologically they offer opportunities to create new and more detailed statistics at different aggregation levels, at lower cost and faster.

2.2. Data storage and access

As collecting, storing, sharing and processing data become increasingly cheaper, the amount of data grows at a staggering speed. All these data are stored in various places and in many different technical solutions,[9] even within NSOs. To satisfy the emerging demand for new information, NSOs must have access to more data and must be able to combine and share these data, thereby keeping data security active. The challenge is to get fast, easy and free access to all relevant data.

Sometimes, data might seem to be held hostage in IT systems, but that is the smallest technological hurdle. The challenge to accessing stored administrative and sensor data is often in addressing the public perception and subsequently gaining acceptance. This is an especially delicate task when the data hold information perceived to be privacy sensitive. Legal access frameworks exist in some countries (such as the Netherlands and Canada) for administrative data that are used for official statistical purposes. The same sorts of legal frameworks exist or are under development in some countries for privately owned sensor data.

Access does not mean that data need to be collected and stored in-house. In many cases, it is enough or even better to be connected to the data. **Data collection through data connection** will be one of the future collection modes, keeping in mind that processes and outputs need to be reproducible, and that time series need to be kept intact whenever possible.

In addition to social acceptability, there are other reasons why data stored elsewhere might not be made available to an NSO: the amount of data may be too large to copy, the data may not be allowed legally to leave the premises, the organization may not yet be ready to send the data (reinforced by the introduction of the *General Data Protection Regulation* in Europe), or only the proportion of data that is needed can be accessed (proportionality aspect of privacy) (e.g., data on business operations in the cloud, public transportation data, ATM transactions and banking data, medical files).

7. Sensor data (self initiated or existing) from mobile devices, wearables or Internet of Things (IoT) type sensors (e.g., optics, RFID, smart meters, connected cars/factories/customers, health care, weather sensors, intelligent buildings, traffic loops, smart surveillance cameras, smart grids…), social media, administrative data (registers and registrations), scanner data, web (scraped) data, etc.
8 Salemink, I. 2019. *Advanced data collection – An outlook to the future*. ISI Conference. STS session 493—Vision on Future Advanced Data Collection for Official Statistics. Kuala Lumpur.
9. SQL databases, data warehouses, flat files, web services, cloud applications, social media, NoSQL databases, semantic networks, etc.

Therefore, solutions need to be found to connect data from NSOs and external parties for statistical research and production, to provide access to external historical raw data, to work with half-fabricates and computation by multiple parties, etc.

To share and use data located outside NSO offices, four patterns can be identified for this microdata linkage:

1. The external data are brought to an NSO to be linked with NSO data.
2. The NSO microdata are brought to an allocated highly secure environment to be linked with other data.
3. Both external and NSO data are kept onsite, and the data are matched by virtual connectivity while being closely managed.
4. The data or algorithm is encrypted and sent to the other data sources for matching, without exposing the inputs provided by other parties (secure multi-party computation).

These patterns and the solutions to store and access data come with capabilities[10] that need further development. **Metadata management**, which is the basis for search-and-find and is a prerequisite to automated privacy-preserving data sharing, is important.

Furthermore, **data virtualization and data abstraction** can simplify data access—regardless of where and how data are stored—by enabling data users to understand information in a way that does not require detailed knowledge of the underlying physical and technical implementation. Retrieving, combining and processing data are made possible without actually having to move or copy the source data. This creates flexibility and faster connections to new data sources, increases data reuse, and connects centralized and distributed data.

A key feature is the ability to discover, manage, disclose and share data sources—regardless of location, volume or structure—and execute data transformations and combinations that meet user needs without bothering the user with techniques. Results are presented in real time, and **data as a service becomes a reality**.

Other relevant capabilities that need to be available are **authentication and authorization** (secure identification of uses and access control), and **classification** and **disclosure control** (on-the-fly-pseudonymization, anonymization, data suppression, perturbation) to preserve the usefulness of the data outputs while protecting confidentiality and privacy.

The development of data service centres for specific economic branches or sector-wide data hubs (e.g., national facilities) with various public and private data complements the NSOs' linking patterns. Data providers can indicate who can use their data and for what purpose by using the same capabilities mentioned before. For all organizations to be successful, efficient and effective, data sharing is necessary, and **data owners must apply the FAIR (findable, accessible, interoperable, reusable) principles to their stored data**. These guidelines will help to improve the findability, accessibility, interoperability and reusability of digital assets[11] by computational systems (machine actionability).

2.3. Computing power and analytical capabilities

IT developments have increased computing and processing power enormously. This increase in performance has made it possible to analyze and use today's huge datasets.

Future datasets need matching between multiple data sources and require knowledge about those sources, their metadata and the quality of the data. Furthermore, the combined datasets have become too voluminous, varied and volatile to be processed and handled with traditional data management processes and tools.

10. See the United Nations Economic Commission for Europe Common Statistical Data Architecture capability model.
11. Wilkinson, M.D., et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* (3). The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object) and infrastructure.

When NSOs and other institutions want to perform joint analysis on each other's datasets, **privacy-preserving analytic**[12] **techniques** allow these privacy-sensitive data from various sources to be analyzed to create new statistics and insights without the risk of divulging confidential microdata. For example, techniques such as secure multi-party computation (SMPC), based on the secret sharing[13] of computation inputs and intermediate results, prevent participants from learning anything about the inputs provided by others.

NSOs are developing proofs of concepts using SMPC and block chain technology. However, their availability and use in official statistics are still in early stages. Currently, no strong development environments appear to exist outside academic research.

These privacy-preserving techniques can be used in combination with the data virtualization approach to simplify data access, regardless of where and how the data are stored, which makes this a powerful combination for setting up joint analytical platforms.

Statisticians traditionally focus their analyses on data that have been collected for statistical purposes. This deductive reasoning puts the idea or hypothesis first, and the analyses are executed to explain, check or validate the idea. Because extensive data are now available that have been collected for reasons other than statistical purposes (and often not under an NSO's control), there is an opportunity to generate new ideas (inductive reasoning). To enable data-driven decision making, evidence-based policy making, and continuous improvement, these two analytical approaches should be seen as complementary,[14] proceeding iteratively and side by side. **Explore *and* confirm** should become **the analytic trend of the future**.

Much how privacy-enhancing techniques are deployed as close to the data owner as possible, a similar trend is seen for analytics and data quality control. Since cleaning up data downstream at NSOs is expensive and not scalable, it becomes important to execute analytics and related data quality processes on the data-gathering devices themselves (i.e., at the edge). This is called **edge analytics**. Similar to edge computing, the trend is to move the analytics and the data quality frameworks to the data instead of moving the data to the centralized analytics and quality frameworks.[15,16]

These examples of a paradigm shift show that we need to work collectively to move forward quickly and develop these methods and approaches so they become standard. The statistics community must work together to establish and implement these standards across the board.

12. For an overview, see the United Nations handbook on privacy-preserving computation techniques: https://marketplace.officialstatistics.org/technologies-and-techniques.
13. Shamir, A. 1979. "How to share a secret." *Communications of the ACM* 22 (11), 612–613.
14. @DiegoKuonen, kuonen@statoo and www.statoo.info.
15. United Nations Economic and Social Council. 2016. *Report of the Global Working Group on Big Data for Official Statistics,* Statistical Commission, 47th session, March 8, 2016.
16. Handle the new in new ways; push computation out (partially). Eurostat. 2018.

3. Challenges and constraints

Thus far, a number of strategic insights and trends have been presented. A number of challenges and constraints have also been pointed out. These will be developed further in this section.

3.1. Data gaps

The increasing demand from society for data-driven, fact-based information causes data gaps in statistical information that cannot be bridged through the traditional survey-based approach.[17] Surveys are too costly and time-consuming, and there is no guarantee that they are always adequately measuring the current reality. For example, the labour market is increasingly dynamic, with new types of employees emerging, such as the self-employed, part-time workers and freelance workers. Currently, the Labour Force Survey questionnaire does not always capture all of these emerging groups. New data sources such as LinkedIn or electronic registers from social security agencies could provide an opportunity to help measure this new reality without having to increase response burden.

Another example of a data gap is policy makers' specific interest in the economic contribution of family businesses,[18] self-employed labour, almost-failed firms and enterprises involved in the Internet economy.[19] These subpopulations can be derived only by combining various data and linking administrative data to statistical units in the Statistical Business Register. In that way, datasets can be enriched, enterprises can be characterized, and subpopulations can be determined and differentiated.

The necessary monitoring of the many SDG indicators will no doubt bring to light many data gaps that conventional data collection methods cannot bridge.

Alternative sources are not necessarily replacing current information; they are filling information gaps. **Secondary administrative and alternative (sensor) data are needed as new sources of information to better assess the current society and to comply with new tasks**.

3.2. Burden to society and response rates

The global decrease in respondents' willingness to participate in social surveys and the aim to reduce the burden of data collection on businesses and citizens require new methods and improvements to data collection strategies. They are also giving rise to a major change in data collection. As other researchers have deliberated,[20,21,22] there is broad agreement on the need to move from a survey-centric model to a model that blends structured survey data with administrative and alternative data sources.

If surveying remains necessary, it should be carried out as **efficiently, easily and least burdensomely** as possible. One of the most viable options, which will be discussed further in Section 4, is to use the reverse surveying process approach ("CAWI-CATI-CAPI") with targeted interviewing teams in NSOs to obtain data from difficult-to-reach target audiences. Using other devices and approach strategies (e.g., calling during certain parts of the day [time slices]) also contributes noticeably.

17. Jarmin, R. 2019. "Evolving measures for an evolving economy: Thoughts on 21st century US economic statistics." *Journal of Economic Perspectives* 33 (1), 165–184.

18. Konen, R. 2017. *Family Businesses in the Netherlands*. Meeting of the Group of Experts on Business Registers. United Nations Economic Commission for Europe, Eurostat, and Organisation for Economic Co-operation and Development. Paris.

19. Oostrom, L., A.N. Walker, B. Staats, M. Slootbeek-van Laar, S. Ortega Azurduy, and B. Rooijkakkers. 2016. *Measuring the Internet Economy in the Netherlands: A Big Data Analysis*. Available at: https://www.cbs.nl/nl-nl/achtergrond/2016/41/measuring-the-internet-economy-in-the-netherlands.

20. National Academies of Sciences, Engineering, and Medicine. 2017. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, D.C.: The National Academies Press. Available at: https://doi.org/10.17226/24652.

21. National Academies of Sciences, Engineering, and Medicine. 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, D.C.: The National Academies Press. Available at: https://doi.org/10.17226/24893.

22. Bean, C. 2016. *Independent Review of UK Economic Statistics*.

3.3. Privacy protection and difficult access to data

The possible use of alternative data sources in conjunction with primary data must be developed under the strict condition that the high level of trust and privacy protection that NSOs have always maintained with the public remain intact. While NSOs continue to work toward using innovative methods to provide more timely, detailed and high-quality statistics and insights, they will have to meet increasingly stringent privacy laws and regulations. NSOs' inviolable commitment not to share any identifiable information they are entrusted with will remain paramount, and concepts such as privacy by design will most certainly become the norm.

In return, free access to data should be the default for NSOs to be able to play their unique role without creating unnecessary costs or burden for citizens and companies. Despite the enabling trends in data storage and data sharing, and the tools available to access and analyze data while preserving privacy and confidentiality, gaining free access to privately held data is not self-evident. Today, NSOs experience easier access to tools, but more difficult access to data.

Applicable privacy-preserving techniques must be developed. The same holds for legal frameworks to establish responsible and ethical facilitated access to administrative and sensor data in such a way that confidentiality and privacy are protected, and social acceptability is obtained. **Data access needs privacy-preserving techniques, legal frameworks and social acceptability**.

3.4. Methodology

The daily work of NSOs is to continually publish data on various socio-, business-economic and demographic concepts. Knowing what needs to be measured to create actionable intelligence about these concepts is essential and becomes more complex given the preferred use of already-available data.

NSOs must establish the extent to which the already-available data cover the concepts, and what additional surveying is needed. This step of conceptualization has always been part of the survey methodology, but with the increasing availability of data, its nature is bound to change considerably. Primary data collection as the starting point will have to evolve into a more holistic approach: **survey methodology becomes data methodology**.

A sound metadata system that describes the data in terms of concepts, techniques, processes and origins is crucial for seeking and finding data, and for future automated classification. This metadata system, which will need to be flexible and expandable, will help statisticians find the appropriate data; understand the data's semantic characteristics; and use the information in data integration, data analyses and other statistical activities.

Because there are multiple survey modes and sources, the complexity of survey designs increases. Complex mixed-mode designs are needed. The real challenge will be implementing "**integration by design**," where the information already available (administrative data, sensor data) is taken into account before making the observation and survey design. This way, instead of enriching surveys afterward with administrative data, the combined administrative and sensor datasets are completed with survey data only when necessary.

The nature and quality of the process of combining data sources will dictate the result. Combining different data sources from multiple survey modes requires an advanced methodology that deals with data source matching and linking issues. Issues could involve linking units without equal source units (individuals, businesses), estimating the correlation between variables occurring in different sources that do not contain overlapping units, correcting for bias, etc. Generic techniques such as probabilistic matching, matching with supervised machine learning, and synthetic matching need to be extended or combined to solve these issues. Combining registers and survey data comes with its own specific challenges when variables occur in multiple sources with different measurement errors. Methods are needed to produce consistent estimates using such approaches.

The technological opportunities and possibilities will determine what data are collected. Producing statistical information from these data will no doubt become more common even though it will be more complex conceptually and methodologically. Future challenges will include validating the results and measuring errors and biases.

Although surveying remains indispensable until proven otherwise, NSOs are steadily losing control of the content of the data that are captured. If necessary, NSOs could partly overcome this issue by developing their own proprietary sensor networks in the future. Given the decreasing costs of sensors and IT technology, this is not too far-fetched.

4. Future data collection

NSOs should aim to use the vast amounts of data that are or that become available in our digital society. These data can be used as inputs for new statistical products, that respond best to users information needs, to supplement existing data acquisition or as replacements for existing survey inputs in the statistical production process.

To be able to do that, it is necessary to develop a new way of looking at data collection and at the official statistical production process as a whole. Instead of acting from specific variables linked to a dedicated collected dataset, NSOs need a paradigm shift. The focus needs to be on reviewing available data, determining what is missing (gaps), and searching for opportunities to design new outputs or to replace part of the existing output.

As a consequence, surveys are likely to become smaller and more tailored, and to be used increasingly for validation purposes. This will no doubt affect the way NSOs conduct these surveys (flash surveys, event-driven surveys, etc.), the survey modes, the devices used and the expectations of the results. Although the nature of surveying will change, the total amount of primary data collection might change either way. The amount of surveying could decrease because of the availability of other data that can be used for a specific statistical product. Conversely, the amount of surveying could increase because it will be possible to create much more useful information on society and the economy.

A new observation strategy has emerged as a result of new data sources and observation techniques becoming available. This strategy includes removing as much of the burden of primary data collection from society as possible. Taking all of the above into account, this **observation strategy** is defined as follows:

> "Statistical output is generated to a maximum extent using non-primary data sources. Searching for available and applicable data sources, data capture modes, and data sharing solutions is an essential part of the collection strategy, as is protecting confidentiality and privacy, with an appreciation for data suppliers and regard for social acceptance."

4.1. Signalling trends and potential future needs

To execute the observation strategy, NSOs must closely monitor social developments in future information requirements; social, economic and demographic policy needs; and national and international social, economic, cultural, technological and ethical trends. This information will translate into future needs for information products. Co-creation with statistical product end users, data owners and NSOs will help to complete the picture and develop a demand-driven portfolio of products based on new data sources. **Collaboration gives information on end-user needs, access to data, infrastructure and knowledge.**

4.2. Tapping into and unlocking new data sources

New data emerge in many forms and from many sources that all come with their own particularities and retrieval challenges. This section will discuss three types of data: administrative data, sensor data (in the broadest sense) and data originating from data platforms (public or private) hosted outside NSOs. Advanced data collection requires new skills, such as data scouting, which relates directly to tapping into and unlocking new data.

4.2.1. Data scouting and collection of administrative data

Although registers and registrations are common, an NSO cannot always easily access the data at its disposal. Firstly, the NSO must know that a data source is available, and, secondly, it must determine whether the data source (combined or not with other data sources) is useful for official statistics. It is at this interchange between hard skills in data handling and soft skills in relation management that Statistics Netherlands (CBS) defined the position of data scout.

A data scout's role is to organize and help acquire and open up new data sources to create new statistics, or to improve or enrich existing statistics. A data scout is the link between internal and external stakeholders. The data scout's portfolio comprises a wide range of tasks that vary between working with content experts to identify the NSO's data requirements; mapping possibilities of new, relevant data sources; evaluating and testing the usability of possible new data sources; working with legal, technical and domain experts to discuss possibilities; building new relationships with relevant partner organizations; negotiating terms and conditions for data use; defining joint business models; and making agreements with data owners.

Challenges that data scouts may encounter include needing to continuously anticipate future possibilities (anticipating possible applications before completely knowing the data), managing long project lead times, setting up joint business models with private corporations, depending on external parties, resolving issues concerning data sharing (safety protection, access, storage, etc.) and navigating various legal issues (privacy, the *General Data Protection Regulation*, collaboration with private partners, etc.).

4.2.2. Sensor data to collect information

Sensor data have become omnipresent in business processes and in daily life. As a result of their potentially high population coverage and use in daily life, Internet-of-Things (IoT) sensors, wearables and mobile devices have become tools that can supplement or replace surveys with automatically generated sensor data.[23]

> **Examples:**
> A well-known application of **IoT sensor data** is the use of satellite imagery, which is a key component for statistical programs on energy transition and agriculture. Satellite images and aerial pictures are used to get a complete and detailed picture of installed solar panels at the regional level. Combining satellite imagery with climate data enables crop and pasture conditions from which vegetation index values are collected to be depicted in near-real time. The results are accurate enough to replace traditional collection methods, eliminating a substantial amount of surveying.
>
> Other examples include using traffic loop data to estimate regional traffic intensity, and using public transit chip data to provide information to municipalities to develop public transit policies. Partnerships between data owners and NSOs offer excellent opportunities for each organization to reach its goals.

The focus of using sensor data that are already being collected by companies (for example weight of crops in harvesting machines) is to recycle the information in an automated way in order to reduce the response burden for the farmer in this case. A thorough analysis of the data also allows to detect further use of the data for users and investigate the potential for new official statistics. Finally, it allows questioning whether current statistics are still sufficiently addressing current user needs. Linking with already existing data is then considered an important final step.

In the same way that smart technologies and smart data (sensor data, social media data) influenced official statistics to create smart statistics, they also made their way into surveys to create smart surveys. Respondents to smart surveys use **smart personal devices**, **wearables and mobile devices** to provide survey data using the **built-in sensors** (e.g., accelerometer, GPS, microphone, camera).

The collected sensor data usually replace or supplement questionnaire data. Therefore, smart surveys go well beyond merely using web-based (online) data collection, which essentially transforms paper questionnaires into electronic questionnaires (this, in itself, is also subject to innovation; see 4.3.2). Instead, smart surveys involve dynamic and continuous interaction with both the respondent and the personal devices, thereby combining self-initiated data with passively collected sensor data.

23. https://www.cbs.nl/en-gb/our-services/innovation.

Survey topics that require in-depth knowledge (e.g., travel locations), that are cognitively burdensome (e.g., consumer expenditure or time use diaries), and that are simply hard to translate into questions (e.g., physical condition) are excellent candidates for smart surveying. Promising statistical topics for exploring smart surveys include budget expenditures, health and lifestyle, time and media use, living and working conditions, mobility, and travel.[24] Apart from using the built-in sensors in mobile devices for data collection, NSOs could develop specific **data collection applications (cross-platform applications) to run on mobile devices**—a popular research topic. For example, these applications could measure time-location data and traffic movements, combined with mode and method of movement. In addition to providing a convenient way for respondents to complete some of the most burdensome surveys, these applications could use notifications to nudge respondents at strategic times and get them to respond. Data collected elsewhere on the device could be used by the application, provided that the respondent consented and the data meet the project requirements. Further to these developments, another category of sensor data may be introduced in the future by **proprietary sensor networks** developed and exploited by NSOs. With these proprietary networks, data collection would be automated, like with sensors, but the data would be produced specifically for statistical data collection.

Of course, the fact that it is possible to collect sensor data (self-initiated or existing) is not enough reason to do so. Self-initiated sensor data require a data collection infrastructure and lead to direct data collection costs. They require a new data collection channel, new processing tools and new skills to expand existing monitoring and analysis tools and to redesign the survey estimation methodology. Other aspects such as data storage, privacy and legislation are also different and important. On the respondent side, the sensor data may still be burdensome and privacy intrusive. To determine the utility of sensor data for official statistics, criteria to support cost-benefit assessments are helpful. Three perspectives may be explored: survey quality cost, sensor quality cost and respondent experience.[25]

4.2.3. Data platforms and other infrastructure as data sources

The data cloud is usually referred to as a storage method or a service. However, it may also be a valuable data source that justifies investing in cloud technology. The cloud can be departmental (e.g., government-wide financial data cloud), data service centres that join businesses with business operations data or international (e.g., the United Nations Global Platform). Also, a collaborative data platform[26] can be valuable as a data source to bring together various scientific and NSO data collections.

NSOs themselves can also work to further expand their role as data hubs and producers of statistics from and for governments, linked to national data strategies. NSOs can be platforms that enable broad cooperation between municipalities, scientific institutes, governmental bodies and businesses. A data ecosystem aims to create an environment in which the cooperation between NSOs and decentralized governments results in clusters of innovative enterprises and institutes and uses the rich data infrastructure that participating parties have to offer. NSOs become facilitators and integrators, using their expertise to create better and fit-for-purpose results through open innovation while maintaining quality, privacy and confidentiality.

4.3. Future surveying

In the foreseeable future, primary data collection will remain an important part of data collection for official statistics. In view of continuously decreasing response rates, initiatives are needed to motivate respondents (e.g., reminder SMS messages; active management based on current, timely and empirical observations; continuous improvement of respondent communications; and targeted group approaches to "nudge" respondents to respond). Because of the importance of surveys, and to validate sensor data, there is a need for substantial research, development and innovation.

4.3.1. Multi-mode: CAWI-first → CATI → CAPI

24. Mussmann, O. and B. Schouten. 2019. *Final Methodological Report Discussing the Use of Mobile Device Sensors in ESS Surveys—Sensor Data for ESS Surveys: A First Inventory*.
25. De Broe, S., G. Snijkers and B. Schouten. 2019. *Sensor data at the heart of innovation in official statistics*. ISI Conference. STS session 493—Vision on Future Advanced Data Collection for Official Statistics. Kuala Lumpur.
26. For example, SURFsara, hosted by ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations). See https://odissei-data.nl (also available in English).

A strategy that uses computer-assisted web interview (CAWI) first is seen as an important step to improve response rates. However, it will not be sufficient in the long term. Survey respondents increasingly expect an electronic self-response mode. Both Statistics Canada and CBS set out to build this option for their respondents, while at the same time replacing a myriad of data collection systems that were becoming increasingly difficult and costly to maintain.

The new system has resulted in approximately 80% of Statistics Canada's surveys now offering an HTML-based, multi-mode-ready questionnaire that can be delivered to a respondent's computer, laptop or other mobile device, and that can be accessed by interviewers at home or in call centres. The remaining 20% of surveys are planned for migration to the new system within the next 24 months. At CBS, responses to business surveys are gathered almost completely through the Internet. For individuals and households, the response rate varies between 50% and 77%, depending on the type of survey and its design.

The electronic questionnaire platform is achieving two goals. The first goal is to provide respondents with their preferred response mode. The second goal is to establish cost savings, since this self-response mode is reducing the number of hours of interviewing required. At Statistics Canada, the estimated annual savings from offering an electronic questionnaire option (not including the census) are CAN$2.9 million so far. CBS reduced its computer-assisted personal interview (CAPI) staff by 50%, from 200 full-time equivalents to 100. This resulted in a yearly salary cost reduction of €135,000. By introducing adaptive survey design and time slices, CBS also anticipates a cost reduction for computer-assisted telephone interviewing (CATI) between 30% and 50%.

4.3.2. CAWI hybrid mode

The proliferation of mobile devices has made them the standard communication tool. Allowing respondents to respond online has resulted in a steady increase in the proportion of online logins by mobile devices.[27] The **CAWI mode becomes hybrid** because of the variety of online devices used to log in (desktop, laptop, tablet and phone). Consequently, to meet present-day requirements, web surveys have to be smartphone and tablet compatible and mobile friendly.[28,29] Survey tools have to be able to automatically scale in a methodologicaly correct way to different screen sizes (e.g., as implemented in the CBS Blaise data collection platform[30]) . Also, these tool have to be able to support CAWI hybrid modes with readability for every question type (single/multiple choice, open-ended, dropdown, slider, grids, ranking, etc.), easy selection, visibility across the page, simplicity of design features (buttons, icons, scrolling) and predictability across devices.[31]

4.3.3. Multi-mode and multi-source data collection

In addition to adopting the multi-mode approach, more NSOs are following the trend of moving from single-source to multi-source statistics. Administrative data have proven to be valuable in combination with data from sample surveys and registrations because they often contain more information than sample surveys can provide. Combining data sources means that new and more detailed statistics can be produced faster (examples in text box below).

Another hybrid approach is using web-scraped information to pre-fill questionnaires to reduce non-response. Automated web scraping involves important ethical and privacy considerations for NSOs. NSOs are advised to develop a directive on web scraping to establish recommendations about using an application programming interface (API) when available, consulting the websites' robots.txt, respecting website controls and protocols, not capturing personal information, being transparent, and addressing other issues. Web scraping could also be explored for gathering information that can support enterprise profiling, financial variable coherence analysis, merger and acquisition event detection, and sentiment indicators for measuring business tendencies based on news analytics.

27. https://www.cbs.nl/en-gb/background/2018/06/mobile-device-login-and-break-off-in-individual-surveys.
28. Lorch, J., and N. Mitchell. 2014. *Why You Need to Make Your Survey Mobile Friendly Now*. Shelton: Survey Sampling International.
29. Antoun, C., J. Katz, J. Argueta, and L. Wang. 2017. "Design heuristics for effective smartphone questionnaires." *Social Science Computer Review* 36 (5), 557–574.
30. CBS Blaise. 2018. *Blaise—Gaining Deeper Understanding (Methodological correct scaling developed in collaboration with the University of Utrecht)*.
31. Bakker, J. 2018. *Designing the Questionnaire of Tomorrow, Current Best Practices and Future Goals*. IBUC Conference, Baltimore.

Interaction and dialogue is now also part of respondents' relation with data. In connecting with millennials and other hard to count populations[32] adding narrative to a survey could be the next step. Other areas for innovating surveying modes are the development of WhatsApp surveys, introducing game elements to increase interest in surveys (gamification), virtual and augmented reality.

Examples in practice:
Census: The Dutch census is a population census for which no additional data collection is needed (0 enumerators). By bringing together all available data sources, combining registers and surveys a virtual census is conducted see for the details and explanation https://youtu.be/SLpDkcyenf0

Mobility: Combining administrative data on vehicle ownership, vehicle characteristics, driver's licences and distances travelled with characteristics of individuals and households, to gather information on traffic and mobility related to trends in society, and to develop more regional data and information on specific population groups.

Consumer Price Index (CPI): Using scanner data to produce the CPI, where the ultimate goal is to replace all food-price primary data collection in the field (in-store collection). With this approach, each field-collected quote is being replaced by an average price for the same or similar product using scanner sales data. The scanner data need to be pre-processed to link the products to the CPI classification. Machine learning can be used for this classification process.

4.3.4. Custom-fit data collection

Adaptive and responsive survey design has received significant interest over the past decade. It might be an effective way to counteract budget pressures caused by declining response rates.[33] Face-to-face observation can be reduced through random selection and through stratified selection of non-respondents eligible for face-to-face follow-up. The latter method could potentially reduce non-response bias. This adaptive survey design assumes that adjusting efforts for relevant population subgroups is either effective in improving survey quality or efficient in reducing survey costs. The key decision to be made is how to divide the population into strata, which is done by using a classification tree. People are divided into groups based on personal characteristics (e.g., demographic and regional characteristics that are known to have a different response distribution than the population).[34] Examples of characteristics are ethnicity, ethnicity of parents, age, income, urban characteristics of the neighbourhood or municipality, education, household type and size, marital status, wealth, gender, and home ownership.

Case prioritization is another adaptive approach to improve sample representativeness by targeting high-priority surveys or cases that belong to domains with lower response rates. In some circumstances, case prioritization might be used to target specific cases for various operational reasons. The objective is to monitor data collection while it is in progress to identify the cases to prioritize. This method uses information available before and during collection to adjust the collection strategy for the remaining in-progress cases. The allocation of interviewer efforts is related to case prioritization; interviewer efforts can be prioritized on cases where they will be most efficient.[35]

32 See https://www.census.gov/library/stories/2019/07 by Schwartz, Z.
33. Chun, A.Y., S.G. Heeringa, and B. Schouten. 2018. "Responsive and adaptive design for survey optimization." *Journal of Official Statistics* 34 (3), 581–597.
34. van Berkel, K., S. van der Doef, and B. Schouten. 2018. "Implementing adaptive survey design with an application to the Dutch Health Survey." *Journal of Official Statistics* (submitted).
35. Laflamme et al. 2016.

4.3.5. Experimenting with new primary data collection methods

Solely focusing on optimizing current primary data collection operations will be insufficient. New data collection methods must be explored to reflect the new reality of a population less interested in completing surveys. The advantages of available technologies must be explored to transform primary data collection operations. A few anticipated feasible examples include the following:[36]

Examples:

Developing a crowdsourcing service: Crowdsourcing involves asking the population to proactively provide information rather than wait to be contacted when selected as a respondent. The risk of such an operation is obvious to a statistician—crowdsourcing data quality is difficult to assess, and quality metrics are nearly impossible. However, crowdsourcing at Statistics Canada has shown interesting results with the introduction of new primary data collection techniques (e.g., GPS locations for a set number of dwellings, the price of cannabis[1]). An alternative, but comparable, approach is using "spontaneous response," a variant of web interviewing where questionnaires are presented at strategic, frequently visited Internet sites where visitors are encouraged to participate in a survey.

Testing cognitive interactive voice response (IVR) technology: Statistics Canada plans to test cognitive IVR technology as an interviewer or respondent support tool. Cognitive IVR is a way for humans to interact with an artificial intelligence platform, such as IBM's Watson or Google Duplex. The research will focus on exploring ways to automate quality control and interviewer feedback to better tailor the tone and approach to each individual respondent. In addition to providing automated feedback to human phone operators, this technology could allow respondents to call a phone number and be interviewed by the cognitive IVR system, just like they would with an interviewer.

36. For a more detailed description, see Dufour, S., G. Bowlby, F. Laflamme, S. Bonhomme, and H. Mullin. 2019. *Modernizing Data Collection in Canada*. ISI Conference, STS session 493—Vision on Future Advanced Data Collection for Official Statistics. Kuala Lumpur.

5. Conclusion and call to action

Implementing advanced data collection capabilities is crucial for national statistical offices (NSOs) to increase their added value for the societies they serve. The datafication of society, cost-efficient storage capacity and a rapid increase in computing performance have opened new opportunities for using sensor data in conjunction with more traditional data sources (surveys and administrative records).

Many areas must be taken into account to bring out the full future potential of NSOs, including new data sources, new collection methods and collection process redesigns. All these options come with certain consequences with respect to methodology, technology, quality, metadata and other standards, confidentiality, privacy, acceptability, and more. The amount of knowledge development required calls for extensive collaboration not only between NSOs, but also with governments, end users, academic institutions, research organizations and private sector companies. This position paper has presented a vision of future advanced data collection and its implications. This last section summarizes the conclusions and presents a call to action for all organizations that can contribute to realizing this vision.

The objective of NSOs is to provide society with information on complex societal and economic phenomena. The information needs to be as actionable as possible so end users can use it to support evidence-based policy making, decision making and opinion forming. Trends in society require information at all aggregation levels (international, national, regional, local). Furthermore, end users clearly benefit from more timely information to react to changes in society more adequately.

To maximize its added value, the statistical sector needs to take these trends seriously and develop means and methods to comply with these user demands. It is obvious that these demands cannot be met using conventional data collection methods.

This paper deals with making available the raw material (data) that is needed to produce the necessary statistical products and services to meet the end-user requirements. Next to primary (survey) data, new types of data (secondary, tertiary and even quaternary) are or will become available, all with their own characteristics. These new data types have been discussed to some extent and can be classified as summarized in the table below.

		Data	
Type	Source	Collected for official statistics	Collected automatically
Primary	Survey *	YES	NO
Secondary	Administrative **	NO	NO
Tertiary	Sensors ***	NO	YES
Quaternary	Proprietary sensors	YES	YES

* CAWI (i.e. hybrid mode including desktop, laptop, tablet, mobile phone), CATI and CAPI.

** Registers (constantly updated files with data on persons or affairs) and registrations (recordings of data)

*** Self initiated and existing sensor data, from mobile devices, wearables and Internet of things sensors, includes;
Social media, scanner data, web (scraped) data,
Smart surveys using smart personal devices (wearables and mobile devices) using the built-in sensors, combining self-initiated data with passively collected sensor data,
Cross platform applications.

The main challenges for the future are making these data available; using and combining them to create official statistics that meet the above-mentioned criteria; and complying with demands on quality, confidentiality, privacy, acceptability and more.

Making these data available to NSOs does not mean that all these data need to be stored and maintained on NSO premises. Accessing data by connecting and linking is the future hybrid data collection method.

It is clear that, before this can become a full-scale reality in the official statistics community, a vast amount of fundamental research and development needs to be done.

This vision paper demonstrates that existing data collection methods need to be redesigned and optimized, and that new data collection methods need to be developed in conjunction with developments in methodology, IT, legal frameworks, etc.

Open innovation within and outside the statistical community, and co-creation with end users, is crucial to proficiently reach this goal.

5.1. Focus areas

The following topics emerge from the vision on future advanced data collection presented in this document. They need further elaboration, research and development.

5.1.1. Methodology

Because of the multiple collection modes and different data sources, complex mixed-mode designs are needed. Preferably, they should integrate already-available data in advance: integration by design.

Linking different data sources and validating data that were not collected specifically for official statistic purposes (administrative and sensor data) require completely new and advanced methodological concepts. Changing from a survey methodology toward a data methodology is key.

Interaction and communication with respondents are fixed values and remain an essential part of data collection for the foreseeable future. Aspects of behavioural science become a substantial part of survey methodology as part of data methodology.

Use case methods to improve timeliness are a must. The usually fixed nature of the non-proprietary raw material (data) means that solutions need to emerge from statistical production methods like event-driven processing (EDP). Statistical methods need to be adapted to introduce EDP and, thus, become more responsive to actual events.

5.1.2. Quality

Timeliness will become an important characteristic of the quality of statistical products and, of course, timeliness might influence the accuracy of statistical information. Accuracy is not a synonym for quality, but it is one of the characteristics that determine the quality of statistical products. As long as the accuracy of information is known and specified to the end user, this need not be a problem. Research into reducing the potential trade-off between timeliness and accuracy would of course be of interest.

As a matter of course, the concept of fit-for-purpose statistical information will never influence the scientific integrity, transparency and independence of the work done by NSOs.

5.1.3. Data access with respect to social acceptability and legal frameworks

Social acceptability is key to gaining access to privately held data. This means that NSOs need to be transparent and able to demonstrate and explain the value proposition to society (public good) and address society's concerns about trust, confidentiality and privacy.

At the same time, legal frameworks need to be developed to make sure that NSOs can use all these new data sources to their maximum extent and to make sure that society can benefit from the added value NSOs can potentially provide.

5.1.4. Data access with respect to technology and methodology

The keywords for future data access are collect, connect and link. Technology to obtain secure data access, in conjunction with the appropriate methodology and algorithms to guarantee privacy and confidentiality, is one of the main technological development areas for the near future. Multi-party computation, privacy-preserving

data sharing (PPDS) and privacy-preserving record linkage (PPRL) are potentially promising technological advancements that need to be further developed into a robust set of methods.

5.2. Call to action

To take advantage of all the potential opportunities associated with future advanced data collection approaches, much development still needs to take place. Joint efforts in open innovation and further co-creation partnerships and collaboration between the statistical sector, universities, the private sector and end users will be necessary

- **Methodology:** Develop new statistical methods or adapt existing ones to deal efficiently with using administrative data and sensor data in areas such as data linkage, data integration and data validation.

- **Collect, connect and link:** Accelerate the development of techniques and technology to facilitate access to new data sources and ensure even greater confidentiality and privacy protection (PPDS, PPRL, multi-party computation, edge computing, etc.).

- **Metadata:** Develop and promote structures and standards to facilitate the evaluation, curation and linkage of survey data, administrative data and proprietary sensor data.

- **Proprietary sensor networks:** Sensor networks dedicated to statistical purposes need to be designed, developed and tested so that standards and guidelines can be developed and efficient deployment can be accelerated.

- **Legal frameworks and social acceptability:** Legal frameworks need to be developed (or existing ones deepened) to establish responsible and ethical access to administrative and sensor data in such a way that confidentiality and privacy are protected and social acceptability is obtained.

Furthermore, showcases need to be developed to demonstrate the possibilities and sustainability of the new methods and techniques in the areas listed above.

5.3. Open innovation and partnerships

Small targeted coalitions and strategic partnerships are needed to work on these items to unleash the added value of future advanced data collection. CBS and Statistics Canada have taken the initiative for this paper and are prepared to coordinate achieving this vision with all interested organizations that are able and willing to contribute.

The authors can be contacted at:

Marcel van der STEEN - m.vandersteen@cbs.nl
Irene SALEMINK – i.salemink@cbs.nl
Stéphane DUFOUR - stephane.dufour@canada.ca