Harvard Data Science Review • Issue 3.4, Fall 2021

Data Science and Official Statistics: Toward a New Data Culture

Stefan Schweinfest¹ Ronald Jansen¹

¹Statistics Division, Department of Economic and Social Affairs, United Nations

Published on: Dec 01, 2021

DOI: <https://doi.org/10.1162/99608f92.c1237762>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

In the digital age, data are generated continuously by many different devices and are being used by many different actors. National statistical offices (NSOs) should benefit from these opportunities to improve data for decision-making. What could be the expanding role for official statistics in this context and how does this relate to emerging disciplines like data science? This article explores some new ideas. In the avalanche of new data, society may need a data steward, and the NSO could take on that role, while paying close attention to the protection of privacy. Data science will become increasingly important for extracting meaningful information from large amounts of data. NSOs will need to hire data scientists and data engineers and will need to train their staff in these fast-developing fields. NSOs will also need to clearly communicate new and experimental data and foster a good understanding of statistics. Collaboration of official statistics with the private sector, academia, and civil society will be the new way of working and the fundamental principles of official statistics may have to apply to all those actors. This article envisions that we are gradually working toward such a new data culture.

Keywords: data science, official statistics, data steward, data privacy, Fundamental Principles of Official Statistics, data innovation

1. Introduction

Data science can be described as extracting meaningful patterns from data using machine learning (ML), artificial intelligence (AI), and visualization methods. The *Cambridge Dictionary* defines it as the use of scientific methods to obtain useful information from computer data, especially large amounts of data (Cambridge University Press, 2021). The Office for National Statistics (ONS) of the United Kingdom describes data science more broadly as applying the tools, methods, and practices of the digital and data age to create new understanding and improve decision-making (Office for National Statistics, 2021a).

ONS created a Data Science Campus in 2017 to apply data science, and build skills, for public good across the United Kingdom and internationally. The goals of the ONS Data Science Campus are to investigate the use of new data sources, including administrative data and big data for public good, and help build data science capability for the benefit of the United Kingdom. A new generation of tools and technologies is being used to exploit the growth and availability of these new data sources and innovative methods to provide rich informed measurement and analyses on the economy, the global environment, and wider society.

Specific applications of data science for official statistics can be found, for example, in the use of satellite imagery to classify and map crops or to estimate the extent of water surfaces. In those instances, the pattern recognition in the images of a specific crop or water surface is learned based on survey sample data, and it is subsequently applied on new images. Mistakes can occur when certain conditions have not been considered in

the training set of images. For example, in the case of water surfaces, the algorithms initially did not work very well in Canada, because the training sample did not include images of frozen water surfaces.

Another example of machine learning as used in a national statistical office (NSO) is the automated classification of products or services. Based on a description of detailed characteristics, a product item (as sold in a retail shop) is classified in a product category (as used by statistical agencies); similarly, a specific service can be classified in a service category. The data science consists here of a combination of natural language processing and machine learning algorithms. Matching of names (of individuals or businesses) or addresses among various data sets is another application.

With the advancements of digital technologies and the widespread use of mobile devices, GPS, and other sensors, large amounts of data are nowadays generated as byproducts of the prime use of the device. Those large data sources (also known as Big Data) can be very useful for policy purposes to determine where populations are moving in emergency situations, how commuting patterns are developing both in time and space around large metropolitan areas, or where tourists like to go. These Big Data sources can also help improve the products and services of NSOs. However, advanced technologies and new skillsets are required, and data science and data engineering¹ will be needed to harness the potential of Big Data for official statistics.

2. Expanding Role of Official Statistics

Gradually, NSOs have started experimenting with the use of new data sources, including Big Data, such as satellite imagery, scanner data, or mobile phone data. Support from other disciplines is needed to tackle the problems inherent in using those kinds of data sources. Use of satellite imagery (provided by space agencies or the private sector) requires machine learning algorithms (developed by space agencies, the private sector, or academia), which are trained on sample survey data to subsequently be applied to target areas. Processing of a large volume of mobile phone data requires appropriate data engineering to build data pipelines that can condense data into meaningful patterns.

The United Nations Statistical Commission (as the apex entity of the global statistical system) created the UN Committee of Experts on Big Data and Data Science for Official Statistics in 2014 to provide strategic vision, direction, and the coordination of a global program on how Big Data and data science could be used to improve statistics and indicators, making them timelier, more frequent, and more disaggregated. Modernization of national statistical systems was seen as essential to remain relevant in a fast-moving data landscape, making statistical operations more cost effective, complementing surveys, and providing more granularities in outputs. Since 2014, progress has been made on methods, tools, and applications of the various Big Data sources for official statistics. Furthermore, several NSOs (such as the United Kingdom, Netherlands, Korea, Rwanda, and Switzerland) created a Data Science Center alongside the NSO, herewith integrating data science and official statistics. In this respect, the Statistical Commission, which has advanced norms and standards for official

statistics over more than 75 years, may need to include other disciplines, like data science, into its work program in years to come.

With demands for new and timelier data ever increasing and with a rapidly changing data ecosystem, NSOs have responded by modernizing and integrating new data sources into regular data production and strengthening their role as coordinators and custodians of data principles and data quality. Over the last year, the new challenges posed by the COVID-19 pandemic have accelerated this transformation and often forced NSOs to reshape their data programs to adapt to the new circumstances. The more extensive use of private sector data during the pandemic has also intensified the need for improved data governance and for developing principles and tools for data privacy protection. In the new data ecosystem, NSOs can broaden their mandate and play the role of data stewards at different levels and with different arrangements, to ensure the efficient utilization of all data sources while safeguarding data quality, confidentiality, and security.

3. Some New Ideas and Considerations

What could be the new role of official statistics in relation to data science? How can the community of official statistics use many different public and private sector data sources, and how can it collaborate effectively with various stakeholder communities? In this section we bring up some ideas and considerations on new ways forward.

3.1. Many More Data Sources Are Generated: Do We Need a Data Steward?

Currently the national statistical system is led by the chief statistician (sometimes also called national statistician or statistician general). The title of director general or president (of a national statistical office) is often used interchangeably with chief statistician (of a country). Some use both titles, where director general signifies the function of manager of the national statistical office and chief statistician indicates the role of gatekeeper of the fundamental principles of official statistics for the whole national statistical system, which includes, among others, making sure that statistical standards are appropriately applied throughout all national agencies that produce official statistics.

Data steward is a newly emerging role, which implies creating order in a complex national ecosystem of data sources, which potentially could be used to the benefit of society. The role of a data steward goes, therefore, beyond the current role of a chief statistician, as it includes giving instructions to all kinds of government and nongovernment institutes regarding the handling of their data.

New Zealand recently developed a data stewardship framework, which enables government to better manage and use the data it holds on behalf of its citizens (New Zealand Government, 2021a). Data stewardship is the careful and responsible creation, collection, management, and use of data. When used securely—protecting privacy and confidentiality—and with public trust, data can provide rich insights to inform decision-making, improve services, and drive innovation. In New Zealand, the chief statistician (head of the NSO) is also the

government chief data steward (New Zealand Government, 2021b). Can this be the future model for many more NSOs?

In Switzerland, the Federal Statistical Office (FSO) is working on an interoperability platform, which makes data and metadata available for use by other government agencies, local governments, the private sector, and academia (Federal Statistical Office, 2021). FSO is the central data steward. In addition, FSO established a Data Science Competence Center, which offers data science as a service across all government agencies. This facility will support the growth and functioning of the interoperability platform.

NSOs are expanding their traditional role. They can play a central role in making more data available for use and providing the tools to do so.

3.2. Privacy Concerns: Protecting Privacy While Keeping Data Relevant.

Public trust in the treatment of personal data is the cornerstone of the work of NSOs, which have always been entrusted to work with sensitive data; for example, population census data, household surveys, or administrative tax records. The public should have full confidence that the privacy and confidentiality of those data are protected, and that data are only used for statistical purposes (and never for any law enforcement activities, for example).

When working with many data sets, privacy concerns could be compounded. Privacy should be protected for any individual data set, as well as any combination of data sets. Significant progress has been made in recent years on the use of privacy preserving techniques (United Nations, 2019). Those techniques combine protection of privacy (through encryption or adding noise or other techniques) with allowing algorithms to use as much detail of the data as possible. This is an area where statisticians, data scientists, and data engineers work closely together.

When considering the type of data necessary for statistical analysis, statistical agencies should always consider whether their goals can be achieved using aggregated nonidentifiable data instead of identifiable data. Furthermore, physical, technological, and organizational provisions should be in place to protect the security and integrity of stored data as well. In the case of mobile phone data, for example, it is in general preferred if the Mobile Network Operators retain full control of their data and process the data within their own internal secure systems to minimize the possibility of breaches in data security.

More data means more privacy protection needed, and privacy-preserving techniques could be the way forward in this respect (Bogdanov et al., 2016; Keerup et al., 2021).

3.3. Statistics, Indicators, and Geospatial Information: Data Integration.

Data science is a discipline that can help in the various efforts to integrate data at geospatial and conceptual levels. Especially, when faced with many data sets with different structures, data engineers and data scientists

can assist in managing the complexity. Close cooperation between data scientists and subject-matter statisticians will be beneficial to derive meaningful insights from complex data integration projects.

Within the official statistics, efforts are underway to have a more holistic approach, in which the economic, social, and environmental dimensions are considered at the same time. For example, the effects of climate change should be studied not only with impact on the environment but also on the economy and on population movements. To enable data integration, concepts and definitions—on the one hand—should be coherent and consistent across the multiple statistical domains. On the other hand, interlinkages need to be created between many data sets across government and with data sets coming from private sector or civil society. Geospatial mapping of statistical information is one way of creating integrated data sets. Data integration is also a major part of current work on the Sustainable Development Goals (SDG) indicators, such as the integration of statistical data and geospatial information (United Nations, 2021a).

3.4. New Job Profiles: Statistician, Data Scientist, Data Engineer, and Data Analyst.

NSOs (and international statistical agencies) have been hiring mostly statisticians for many decades. Only recently has more attention been paid to the computer science skills of the statisticians. New job profiles are now developed to specifically hire data scientists, data engineers, and data analysts. In the United Nations Data Strategy, data scientists solve emerging and complex organizational problems in a data-driven way (United Nations, 2020). To do so, they design and construct data modeling and data production processes using prototypes, algorithms, predictive models, and custom analysis; a data engineer is a colleague whose primary job responsibilities involve preparing data for analytical or operational uses. The specific tasks can vary, but typically include building data pipelines to pull together information from different source systems; integrating, consolidating, and cleansing data; and structuring it for use in analytics applications; finally, data analysts examine large data sets to identify trends, develop charts, and create visual presentations to help colleagues in program, policy, and operations make evidence-based and data-driven decisions.

Subject-matter statisticians approach the compilation of statistics from the angle of established concepts and definitions, standard classifications, and a quality assurance framework. For example, in the subject-matter domain of tourism statistics there are specific definitions for usual residence, visitor, or a trip. By contrast, a data scientist is first of all interested in compiling insights from one or more data sources, where a conceptual framework may be present in an implicit way, but where standards have mostly not (yet) been established.

For chief statisticians the question is to determine the right mix of statisticians and any of the new categories of jobs. How many data scientists or data engineers would an NSO need? Do we retrain some of the statisticians or hire new data scientists? Statisticians have proven to be very good in developing and maintaining quality assurance frameworks. High-quality and trusted statistics are crucial for the NSO to stay relevant. On the other

hand, as explained earlier, relevant data need to be delivered frequently, in a timely manner, and with much detail. Data scientists are valuable in delivering fast and granular data.

3.5. New Training: Continuous Learning.

Staff of NSOs need to stay abreast of new developments in official statistics and related disciplines, such as data science and data engineering. At the United Nations Statistics Division (UNSD), for example, we are developing an IT strategy that will require a much closer working relationship between IT developers and subject-matter business owners through the application of a DevOps approach and that will include moving most applications to the Cloud. This DevOps approach consists of empowering staff members to acquire not only technical capabilities, but also process and office culture capabilities, while consistently measuring progress in all those areas. Part of the office culture should be continuous learning. In this example, we would require all UNSD staff to understand the basics of the DevOps approach and the fundamentals of how the Cloud works. More generally, statistical offices may need to pay more attention to train staff continuously in emerging fields, like data science.

3.6. Communication of Uncertainty: Experimental Data.

By applying data science and by compiling faster indicators, the NSOs will have to carefully explain to its users (government, business, and the public at large) what the limitations are of the disseminated data. For example, ONS has released many experimental data and analysis on economic activity and social change in the United Kingdom to meet the demand for fast indicators during the time of the COVID-19 pandemic (Office for National Statistics, 2021b). These faster indicators are created using rapid response surveys, novel data sources, and experimental methods. ONS indicated the caveats in releasing these data as follows: “These research outputs are part of the Faster indicators of UK economic activity project and are not official statistics. The indicators are not yet fully developed. We make these data available at an early stage to invite feedback and commentary on their further development. [...], we provide an early picture of activity that supplements official economic statistics and may aid economic and monetary policymakers and analysts in interpreting the economic situation.”

Whereas government, business, and researchers will ask from the NSOs to release data as early as possible, NSOs will need to assure a minimum level of quality and will need to clearly explain the limitations of experimental data. Revising early indicators and benchmarking official statistics will become even more important than before.

A related responsibility for the NSO is to educate the user. It is not sufficient for the NSO to disseminate high-quality data with detailed meta-data. Clear and unambiguous communication of the data is becoming increasingly important. What story is the data telling? Data visualization can help bring the correct message to the public (or to the policy maker). As is explicitly described in the Fundamental Principles of Official Statistics, the statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics

(United Nations, 2014). With misinformation being a major problem in recent years, the NSOs can make a significant contribution to provide independent and trustworthy data in the information system of a democratic society.

3.7. Could We Have Fundamental Principles for the Larger Data Landscape?

For the community of official statistics, it is essential that the public has trust in the integrity of official statistical systems. Confidence in statistics depends to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members. In this context, professional independence and accountability of statistical agencies are crucial. The Statistical Commission adopted 10 Fundamental Principles of Official Statistics in 1994 to assure independence and accountability (United Nations, 2014). These fundamental principles begin by stating that official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy, and the public with data about the economic, demographic, social, and environmental situation. They further state, among others, that to facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods, and procedures of the statistics; and that data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records [or other]. Statistical agencies are to choose the source with regard to quality, timeliness, costs, and the burden on respondents. In 2014, the General Assembly of the United Nations adopted a resolution supporting these Fundamental Principles of Official Statistics, underlining the importance of trustworthy statistics for every society.

With the expansion of the landscape of data, which are being used for policy purposes, should the fundamental principles cover those new data developments as well? In other words, if an NSO collaborates with the private sector, civil society, or other parts of government, should those parties also comply with the fundamental principles, at least within the context of the collaboration agreement? Should we think more boldly about fundamental principles for use of data in public policy (Jansen et al., 2021)?

NSOs have used the Fundamental Principles of Official Statistics to develop the national statistics law or a national code of practice. Those instruments allow follow-up if practitioners do not comply with the principles. What would this mean for an expanded version of the fundamental principles?

With the emerging function for the NSO as national data steward, we do have to consider what this means for a wider application of the fundamental principles.

3.8. High-Level Political Attention to Data: UN World Data Forum.

In recent years world leaders have paid much more attention to the importance of reliable and relevant data than in the past. Misinformation has been called a pandemic. The remedy would be an independent and high-quality source of data, as this has been provided by the NSOs for decades. The demand in general for more

data, and specifically for timelier, more frequent, and more granular data, makes it necessary for NSOs to collaborate with many players in a new data landscape.

UNSD has organized the UN World Data Forum since 2017, in which it brings together all stakeholder communities with a focus on improving the availability and use of data for the monitoring and implementation of the 2030 Agenda for Sustainable Development. Such a large event also draws high-level political attention and could be used to launch some new ideas about the future direction of official statistics, especially in relation to developments in the field of data science.

The 3rd UN World Data Forum took place in Bern, Switzerland, in October 2021 with an outcome document about the Bern Data Compact for a Decade of Action on the Sustainable Development Goals (United Nations, 2021b). In its call to action, the two main points are to (1) develop data capacity, where we strengthen national institutional capacity to continue to modernize national data systems and generate public data that are fit for purpose, open, interoperable, and nationally coordinated, and empower policymakers, planners, and decision-takers to understand these data and use them effectively; and to (2) establish data partnerships, where we continue to engage in public–public and public–private cooperation for the use of administrative records and new and innovative data sources by removing barriers to these sources, while ensuring full respect for privacy and confidentiality.

4. Toward a New Data Culture

In conclusion, data has gained a more prominent position in society, steering decisions at many levels, from the temperature and electricity consumption in your house to real-time information on public transportation to economic forecasts and spending space on the national budget. Within the digital age, NSOs face new opportunities and challenges to use many more data sources and collaborate with a broad group of stakeholders.

In a new data culture, the NSO would collaborate with a broad range of partners from the public and private sectors to produce trustworthy data for better decision-making by government, business, and the public at large. Moreover, the NSO and partners would educate society about how to correctly interpret the disseminated data; among others, through the use of data visualizations and storytelling. Here are a few points that we think are important to position the NSOs in this new data culture:

- Among a wide range of data providers and data users, the NSO may need to take the helm as the national data steward.
- In conjunction with the role as data steward, the NSO will need to revisit national data governance, which could lead to changes in the national statistics law. At the international level, rethinking of the fundamental principles of official statistics could be necessary, or maybe a supplement to the principles for the wider data community.
- With increasing use of additional data sources, the protection of privacy should be systematically evaluated, and privacy preserving techniques may prove very useful.

- Data integration is necessary for the delivery of a holistic approach to integrated public policies. This can be achieved through an integration of statistical data and geospatial information.
- NSOs should accompany the early release of experimental data with clear communication on the purpose and limitations of the data.
- Last but not least, staff of NSOs should receive continuous training in new methods and new technologies, including data science; and NSOs should gradually expand its labor force with data scientists, data engineers, and data analysts.

Disclosure Statement

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations.

Contact

Corresponding author (e-mail): jansen1@un.org.

References

- Bogdanov, D., Kamm, L., Kubo, B., Rebane, R., Sokk, V., & Talviste, R. (2016). Students and taxes: A privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies*, 2016(3), 117–135. <https://doi.org/10.1515/popets-2016-0019>
- Cambridge University Press. (2021). *Cambridge dictionary*. <https://dictionary.cambridge.org/us/>
- Federal Statistical Office. (2021). *Strategy for open government data in Switzerland 2019–2023*. <https://www.bfs.admin.ch/bfs/en/home/services/ogd/strategy.html>
- Jansen R., Kovacs K., Esko S., Saluveer E., Söstra K., Bengtsson L., Li T., Adewole W. A., Nester J., Arai A., & Magpantay, E. (2021). Guiding principles to maintain public trust in the use of mobile operator data for policy purposes. *Data & Policy*, 3, Article e24. <https://doi.org/10.1017/dap.2021.21>
- Keerup, K., Bogdanov, D., Kubo, B., & Auran, P. G. (2021). Privacy-preserving analytics, processing and data management. In C. Södergård, T. Mildorf, E. Habyarimana, A. J. Berre, J. A. Fernandes, & C. Zinke-Wehlmann (Eds.), *Big data in bioeconomy* (pp. 157–168). Springer. https://doi.org/10.1007/978-3-030-71069-9_12
- New Zealand Government. (2021a). *A data stewardship framework for NZ*. <https://data.govt.nz/toolkit/data-stewardship/a-data-stewardship-framework-for-nz/>

New Zealand Government. (2021b). *Government Chief Data Steward (GCDS)*.
<https://data.govt.nz/leadership/gcnds/>

Office for National Statistics. (2021a) *Data science for public good*.
<https://datasciencecampus.ons.gov.uk/about-us/>

Office for National Statistics. (2021b). *Economic activity and social change in the UK, real-time indicators: 2 September 2021*. <https://tinyurl.com/59yt2s44>

United Nations. (2014). 68/261. *Fundamental Principles of Official Statistics*.
<https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf>

United Nations. (2019). *UN handbook on privacy preserving techniques*. <https://unstats.un.org/bigdata/task-teams/privacy>

United Nations. (2020). *Data strategy of the secretary-general for action by everyone, everywhere*.
https://www.un.org/en/content/datastrategy/images/pdf/UN_SG_Data-Strategy.pdf

United Nations. (2021a) *Inter-Agency and Expert Group on the Sustainable Development Goal Indicators (IAEG-SDGs) Working Group on Geospatial Information*. <http://ggim.un.org/UNGGIM-wg6/>

United Nations. (2021b). *United Nations World Data Forum*.
<https://unstats.un.org/unsd/undataforum/index.html>

©2021 Stefan Schweinfest and Ronald Jansen. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. In the United Nations Data Strategy, a data engineer is responsible for supporting everyone with data preparation, speeding up the creation of curated trusted data pipelines and their integration, whereas a data scientist is responsible for extracting deep insights from data using complex models, employing statistics, algorithms, AI, and visualization methods (United Nations, 2020). [↵](#)