Big Data Strategies for Official Statistics

Peter STRUIJS
Coordinator ESSnet Big Data, CBS, Netherlands
Sofie DE BROE
Scientific Director, Center for Big Data Statistics, CBS, Netherlands

Abstract. The environment of National Statistical Institutes (NSIs) is rapidly changing in many respects. The emergence of new data sources provides a number of opportunities for official statistics. At the same time this creates challenges, including how to deal with quality issues when for example developing so-called experimental statistics and turning them into official statistics. As many NSIs have started using big data for statistics, the need for a strategic approach has become increasingly clear. The paper describes and assesses strategic options and explains the big data strategy of Statistics Netherlands, which is now being implemented in the Dutch Center for Big Data Statistics. In essence, big data strategies are about positioning NSIs in the changing environment. The paper identifies the true game-changers for official statistics and formulates the associated strategic questions. Their answers depend on where the value added of official statistics is sought, which is to no small extent related to quality considerations associated with the use of new data sources. New approaches may be called for. In any event, the role of NSIs is bound to change. The traditional role of quasi monopolistic provider of statistics on the many facets of society will erode through the rise of competition. However, the institutional and professional foundation of NSIs may also be exploited for assuming new roles. Ideally this will result in a society that is better informed about relevant phenomena and better equipped to counter tendencies where the value of facts is discredited.

1. Introduction

Big data matters. It creates a number of opportunities for official statistics, but also challenges. Institutes concerned with official statistics have taken up the subject in different ways and with different speed. The UN Economic Commission for Europe was quick in identifying the need for action, resulting in, among other things, an early quality framework for big data [10]. At world level the Global Working Group on Big Data for Official Statistics was created, which looked for example at data access and partnerships [5]. At EU level, the Directors-General of the National Statistical Institutes (NSIs) adopted the Scheveningen Memorandum on Big Data and Official

Statistics [3], which initiated the Big Data Action Plan and Roadmap [4] of the European Statistical System (ESS). The ESSnet on Big Data[1] was created in this context. In order to take up the challenges and accelerate innovation, Statistics Netherlands created the Center for Big Data Statistics (CBDS).

At a more fundamental level, the rise of big data and related changes in the environment of official statistics necessitate a strategic rethinking of the role and position of official statistics in society. This includes a rethinking at the institutional level. This paper describes the strategic thinking at Statistics Netherlands, the options encountered and choices made, and the resulting role of the CBDS. One of the main current challenges discussed is to turn experimental results into official statistics.

2. Changes in the Environment

What are the real game-changers for official statistics? Surely many factors are relevant [6], and most factors are related, but these seem to stand out:

- The *datafication* of society. Individuals, organisations and non-living objects leave a rapidly increasing amount of digital traces. Sensors are everywhere, the Internet of Things (IoT) is rapidly growing. In fact, nowadays it seems that no movement, no action or transaction, no change can take place without somehow, somewhere, creating data.
- Changes in the *distribution of data* in society. The main repositories of data used to be registers of government organisations and the internal administrations of businesses, accessible only to those concerned. Nowadays, a huge amount of data is publicly accessible, in the form of Open Data or otherwise, such as public tweets and information on websites. At the same time, an increasing amount of data accumulates with private organisations, in particular on the behaviour of their clients.
- *Competition* from other organisations. Used to their (near-) monopoly, NSIs are now faced with increased competition from private organisations as well as researchers, who also produce statistical information on phenomena in society. Mobile phone companies use their data to tailor statistics to their customers' needs, businesses use public social media messages to compile and sell real-time sentiment indicators to stock-trading companies, researchers produce inflation estimates for countries whose official statistics cannot be trusted.

[1] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page

- *Changing attitudes* in society. Respondents are ever more reluctant to make the effort to supply data, and privacy concerns are becoming more prominent. At the same time, budget suppliers expect NSIs to do more with less.
- The *dynamics* of the information society. NSIs are used to a stable set of available data sources and a predictable environment. New data sources are increasingly volatile, attitudes in society – such as on privacy matters – are increasingly hard to predict, and the dynamics of technology is famous.

3. Strategic Questions

In view of the changing environment, there are several fundamental strategic questions:

- Where do we want to *position* official statistics and NSIs in society, given the changing playing field? How do we see our role in society in the future? In particular:
    - What *demand for information* do we want to serve and what should we leave to others?
    - Are there any *new roles* we want to assume (or old ones to shed)?
    - What do we want to be our *distinguishing features* compared to other organisations dealing with data, in particular competitors?
- Into what direction do we want our *output* of official statistics to evolve? In particular:
    - How can we best *exploit* new data sources and deal with challenges such as their volatility and fundamental methodological issues?
    - How do we intend to deal with issues of *data access and privacy*?
    - How do new outputs *relate to the existing programme* of statistics?
- What would be a viable and sustainable *business model*? In particular:
    - How do we secure the *financial resources* needed for the realisation of our ambitions?
    - What should be our *market approach* and ambitions?
- In what ways do we intend to *collaborate* with other organisations, what should be our place in the data ecosystem? In particular:
    - In what direction do we want our relationship with other *government* organisations to evolve?
    - What should our relationship with *research* institutes and academia be?
    - What kind of relationship do we want to have with *private organisations*, such as those holding relevant data and IT companies?

There are also many strategic questions that are of a more derived nature, for instance regarding the IT, HR and communication strategy. These questions are also important, but they are not in the scope of this paper.

4. First Principles

For Statistics Netherlands, the strategy depends on its mission. The current mission is "to publish reliable and coherent statistical information which responds to the needs of Dutch society". The responsibilities include the compilation of figures for European statistics. The mission is aimed at enabling society to debate social issues "on the basis of reliable statistical information", as the institution's website states[2].

If the near-monopolistic situation would hold – i.e., Statistics Netherlands would continue producing virtually all the statistical information concerned – the emergence of big data would just increase the possibilities to better respond to the needs of society. However, with big data, competitive statistical information is also emerging. Could this replace the need for Statistics Netherlands to compile similar information?

The simple and short – and slightly provocative – answer is: in principle, yes. However, although there is competitive statistical information on offer which apparently fulfils some needs – some businesses even thrive on this – it is good to keep in mind the value added of official statistics:

- The *quality standards* of official statistics are quite high. Much statistical information produced by others that is publicly available is of lower quality.
- The quality of official statistics is *demonstrable*. There is full transparency of methods and assumptions: Users can verify that validated methodology is applied professionally. For much information available for free on the internet this is not the case.
- Reliable information for debate on social issues requires professional *independence* or submission to independent professional judgement.
- An important principle of official statistics is that the information becomes public to *everyone at the same moment*, for the same price (generally for free). This guarantees a level playing field for all involved in debating social issues.
- The value added of official statistics also lies in their *coherence and comparability*. The bits and pieces of statistical information produced by others do not come close to a coherent statistical programme. NSIs use standards and can link data to comprehensive administrative data sources.

[2] https://www.cbs.nl/en-gb/about-us/organisation

Clearly, for the foreseeable future there is no need to fear competition. There is a continuing need for benchmark statistics. These are indispensable for the validation of statistics based on new data sources, and may be combined with them. However, one needs to recognize and accept that there may be cases where statistics produced by others do fulfil certain needs of society. In such cases the resources can be used for other statistics.

In this context, there might be an opportunity for assuming *a new role*: to validate statistics produced by others. For Statistics Netherlands, this would be in line with its mission, especially since there is an increasing mistrust of publicly available information, leading to "fact-free" discussions [2]. Validation by a trusted authority could then be most helpful. Validation could be done on request, possibly for a fee, or on the initiative of Statistics Netherlands, if considered necessary for guaranteeing that debates on social issues can take place on the basis of reliable information [9].

5. The Role of Opportunities

NSIs are faced with an optimisation problem: What is the optimal statistical output, given the information needs, the available budget, the available data sources and response burden conditions? As long as all conditions are stable, this problem can be tackled in a formalised, (multi)annual cycle in which users are consulted and the output programme established. Then the output changes only gradually, and all activities are essentially output-driven.

In the era of datafication, the conditions are not stable. The proliferation of potential data sources gives rise to enormous statistical output opportunities. In the changing environment new opportunities for funding arise, which is most welcome, given the trend of the budget provided by the central government in many countries. As conditions are not stable, a new approach is called for, where opportunities may be leading. This would complement the more formal annual cycle.

Before describing such a new approach, let us first articulate the statistical output opportunities:

- *New statistics*, describing social phenomena not measured earlier. New information sources make this possible.
- *More detailed statistics*, for instance more points in time, more regional detail, or use of lower levels of other classifications used. This could in particular help solve the problem of the limited resolution of sample surveys.
- *Increased timeliness* of statistics. This can take several forms, such as reduction of the time-to-market, early indicators, nowcasts, or even real-time indicators. Data from new information sources may correlate with and foretell the variable of interest.

- *Increased quality* of statistics, in particular a higher accuracy and lower bias. Coherence and comparability might also benefit. This is especially a possibility if data from new sources is combined with the information normally used.

Taken together, exploring and exploiting new data sources and combining them with already existing sources and statistics will make it possible to compile timely, comprehensive information on social phenomena relevant to users. Thus the phenomena will take centre stage. But there are also opportunities for other stakeholders of NSIs than data users:

- Same statistics with a *lower administrative burden*. Replacement of survey information with data from new sources would have this effect.
- Same statistics with *lower costs*. However, such cases typically require substantial investments and a long transition time.

It is far from obvious how to exploit the new data sources to realise the output opportunities. This requires exploration, new methods, modelling, combination with other sources, etc., and access cannot be taken for granted.

6. A Possible Approach

Taking the foregoing into account, what would an opportunity oriented approach look like? It would have several elements. Obviously, new funding has to be found. Producing output for a fee is certainly one possibility, provided the usual standards are applied, such as making the results available to all members of society at the same time. However, investments – in time, money and effort – have to be made before returns are realised. In fact, some venture capital would be needed. In some cases innovation funds may be used, from the Dutch government (at all levels), Horizon 2020, Eurostat grants (such as for ESSnets) etc., although writing tenders in itself also requires some minor investment.

Another element is *developing partnerships* wherever potential win-win cases can be identified. These can be of a more strategic or a more operational nature [8]. A buzzword of the big data era is the “data ecosystem”. Many organisations face similar challenges; the climate for collaboration seems to be favourable. Plausible partners are those holding data, knowledge organisations, commercial technology providers, potential data users such as government organisations at all levels, international statistical organisations such as other NSIs, and stakeholders in general. Interestingly, the subject of collaboration does not always have to be statistical output. In fact, NSIs could consider assuming new roles, as is explained later for the CBDS.

Still another necessary element of the solution is an *environment for experimentation* with new data, new methods, new technology, new processes, beta products, etc., allowing for exploration and trial and error. This has to be complemented by a capability to do research in order to develop a body of knowledge in a more profound and systematic way. An associated challenge is to link this in an effective way to the existing organisation (of the NSI), especially since a classic problem associated with R&D is the transition from experimental results to regular production.

In this way, the conditions are in place for grasping opportunities and significantly enlarging and improving the capability to respond to the statistical information needs of society. In fact, this implies moving from optimizing the output under conditions of available data sources, budget provided and formalised demand, to maximizing the output, considering all potential new data sources, possibilities to generate funding and opportunities to collaborate with others. This is a *growth proposition*.

However, one element has to be added: *the interaction with (potential) data users*. Within the limits of the resources available at any time, choices have to be made continuously concerning the partnerships to be explored with priority, new data sources to be investigated first, tenders to be written, research priorities, and last but not least: the output to be developed and taken into production. These choices have to be made in view of society's demand for statistical information, but this is in flux. Therefore, some sort of interaction with data users has to be organised, without allowing this to slow down decision making or business.

7. The Center for Big Data Statistics

Statistics Netherlands has created the CBDS to power the implementation of the strategy. It embodies an opportunity oriented approach, including the following elements:

- In line with its name, the CBDS develops *statistics that make use of big data*. This takes place without help from others, or by making use of its network of partners. Moreover, the CBDS is able to provide information as a service through external funding and work for third parties.
- Statistics Netherlands aspires to become *the data hub of the Dutch government* at all levels. Contrary to most government organisations, dealing with data is its core business. This may comprise, among other things, dissemination services for government organisations, knowledge services including analyses, and data linkage services.
- There is a need to develop *data platforms*, thereby enabling the sharing of data from and with partners in a way that maximises the utility of the data while providing all necessary

safeguards concerning security, confidentiality and privacy. This builds on the reputation of Statistics Netherlands as a "trusted partner".

- The CBDS plays a central role in getting *access to new data sources*. This involves so-called data scouting (identifying promising new data sources) and the development of partnerships to secure access. In this context a legal strategy is also being developed.
- The innovation role of the CBDS is based on the further *development of a scientific knowledge base*, for which it has a scientific director. Methodological challenges include dealing with selectivity, developing modelling approaches, solving data linkage challenges, realising visualisations, dealing with volatile data sources, and making optimal use of data that does not directly measure but only correlates with phenomena of interest [1].
- A related, new role of Statistics Netherlands is *offering training on big data statistics*, among others to Dutch government staff. Those educated at Statistics Netherlands in big data research may be expected to facilitate the expansion of the network, collaboration and the development of the market for information and services of Statistics Netherlands.

A related new role is the development of Urban Data Centers (UDCs). As this rapidly expanding activity area is broader than big data, requires management focus and can be seen as a distinct engine of income, this has not been made part of the CBDS itself but is a separate programme.

8. From Experimental Statistics to Official Statistics

The requirements which make official statistics so reliable are too stringent for innovative products. For this reason, the CBDS makes use of the innovation site of Statistics Netherlands, in which experimental products can be presented to the users. Reactions are solicited. However, it is a huge challenge to make the transition from successful research to regular statistics. Much has been learned from the use of a new source, road sensor data, for new statistical output, traffic intensity statistics [7]. Lessons learned include the following:

- Concerning *data sources*, implementation requires continuous quality monitoring of the collected data. Metadata should be (made) available. The source has to be stable enough in contents and access that future output is guaranteed. The interpretation of the definition of variables should be clear and consistent. The relationship between the population comprised in the source and the target population needs to be made explicit if possible. Measurement error needs to be assessed for every source being used in the analysis.
- Concerning *statistical methods*, the method used should be justified in itself and in comparison with alternatives. Standard, proven methods are to be preferred. In case of a new method, a validation assessment by means of a peer review is required.

- Concerning the *statistical production process*, domain specialists need to be involved from the start, for contents and acceptance reasons. New tooling or IT infrastructure may be needed and should be anticipated, especially if the processes for regular production have a different scale or character. Ideally, the migration process should be standardized, and in any case be transparent.
- Concerning the *statistical output*, the results should be reproducible based on the same or similar data at different time intervals by different researchers. Assumptions should be regularly re-tested as part of the implementation process. Validation (and cross-validation) of the results should be undertaken through comparative analysis and disclaimers of the initial experimental product addressed. Users feedback on relevance and usability of the experimental product must be heeded, and legal and ethical issues addressed at a very early stage.

9. Conclusion

In the vision presented, the main strategic questions have been answered for Statistics Netherlands. An effective strategy may contain the following elements:

- In addition to maintaining a more or less stable statistical output programme, including benchmark statistics, take an active, maximising approach to the new output opportunities offered by new data sources, including the dissemination of experimental products.
- Position the NSI in the data ecosystem as a partner actively open to collaboration opportunities, developing a wide network.
- Build the corresponding business model on generating sources of income, developing new or newish roles such as being the government data hub, exploring the possibilities of data platforms, and offering big data courses.
- Make getting access to new data sources a priority, by the use of data scouting and developing mutually beneficial partnerships with data holders, and applying a smart legal strategy.
- Place all activities in the context of augmenting the response to the information needs of society by actively seeking interaction with data users, thereby promoting user empowerment.

Such a strategy would make use of the strengths of NSIs and would result in more, better and faster statistics and services that are in demand by society.

10. References

[1] Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics 31 (2), pp. 249-262. Available at:

http://www.pietdaas.nl/beta/pubs/pubs/jos-2015-0016.pdf (Accessed: 6 September 2018).

[2] Davies, W. (2017). How statistics lost their power – and why we should fear what comes next, The Guardian, 19 January 2017. Available at:

https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

(Accessed: 6 September 2018).

[3] DGINS (2013), Scheveningen Memorandum on Big Data and Official Statistics. Available at:

http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

(Accessed: 6 September 2018).

[4] Eurostat (2014). Big data strategy document. Available at:

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=

0ahUKEwjE6uLTpJHbAhVOUlAKHWR8ChYQFghCMAI&url=https%3A%2F%2Fec.europa.eu

%2Feurostat%2Fcros%2Fsystem%2Ffiles%2Fitem_6_tf_big_data_big_data_strategy_shortened_v2

.docx&usg=AOvVaw0Uj9QCSm7vmAo4jg9Z5REKG (Accessed: 6 September 2018).

[5] GWG (2015). Good Practices for Access and Partnerships. Deliverables 1 and 4 of the Task Team on Access and Partnership of the Global Working Group on Big Data for Official Statistics, 12 October 2015. Available at:

http://unstats.un.org/unsd/trade/events/2015/abudhabi/gwg/GWG%202015%20-

%20item%202%20(ii)%20-

%20Good%20practices%20for%20data%20access%20and%20partnerships.pdf

(Accessed: 6 September 2018).

[6] Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. Statistical Journal of the IAOS 31 (2015), pp 471-481. Available at:

https://www.researchgate.net/publication/282421109_The_opportunities_challenges_and_risks_of_

big_data_for_official_statistics (Accessed: 6 September 2018).

[7] Puts, M., Tennekes, M., Daas, P.J.H., de Blois, C. (2016). Using huge amounts of road sensor data for official statistics. Paper for the European Conference on Quality in Official Statistics 2016, Madrid, Spain. Available at:

http://www.pietdaas.nl/beta/pubs/pubs/q2016Final00177.pdf (Accessed: 6 September 2018).

[8] Robin, N., Klein, T., Jütting, J. (2016). Public-Private Partnerships for Statistics, lessons learned, future steps. PARIS21 Discussion Paper No. 8, 29 February. Available at:

http://www.oecd-ilibrary.org/development/public-private-partnerships-for-statistics-lessons-

learned-future-steps_5jm3nqp1g8wf-en (Accessed: 6 September 2018).

[9] Struijs, P. and Daas, P. (2014). Quality approaches to big data in official statistics. Paper presented at the European Conference on Quality in Official Statistics (Q2014), Vienna, Austria. Available at:

http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf

(Accessed: 6 September 2018).

[10] UNECE (2014). A suggested framework for the quality of big data. Big Data Quality Task Team, December. Available at:

https://ec.europa.eu/eurostat/cros/system/files/Task%20Team%20Big%20Data%20Quality%20Framework_937_unblinded_v1.pdf (Accessed: 6 September 2018).