



# Economic and Social Council

Distr.: General  
4 April 2017

Original: English

---

## Economic Commission for Europe

### Conference of European Statisticians

#### Sixty-fifth plenary session

Geneva, 19-21 June 2017

Item 4 of the provisional agenda

#### The next generation of statisticians and data scientists

### Skills for the new generation of statisticians

#### Note by Statistics Finland and Eurostat

#### *Summary*


This document analyses the competence profile of official statisticians with a particular focus on new data science competences. Modernization of official statistics will depend on the capability to incorporate new data sources and benefit from “disruptive technologies”. This will require new capabilities, skills and competences that may not be part of the traditional skill set of official statisticians.

The document is presented to the Conference of European Statisticians’ seminar on “The next generation of statisticians and data scientists” for discussion.

GE.17-05381(E)



\* 1 7 0 5 3 8 1 \*

Please recycle 



## I. Introduction

1. The capability to incorporate new data sources and to benefit from disruptive technologies will be at the core of the modernization of official statistics. These technologies, for example smart meters, web technologies and user experience platforms, will require new types of skills and competences that were not part of the traditional skill sets of official statisticians.

2. There are numerous examples of new data sources that have potential in this sense: administrative records and registers, as well as large digital data sources, such as road sensors, scanner data or Internet-based data. These large digital data sources are known best as big data. Potential new technologies, on the other hand, range from web-scraping algorithms to linked data opportunities and from multi-mode data collection to combining survey and administrative data. The list can be continued even further.

3. By embracing new data sources and technologies, National Statistical Offices (NSOs) can produce faster, more accurate and more comprehensive statistics adapted to understanding of increasingly complex and rapidly changing global phenomena. In short, new data sources are a way to meet users' needs better. This must be done in a way that does not jeopardise the recognised robustness and quality of official statistics, but that will strengthen our competitive asset in the rapidly changing information ecosystem.

4. One key factor in meeting these challenges is the development and building of the necessary skills and competences. In addition, statistical organizations will have to create favourable conditions for new production methods and using data science skills with success. This entails, for example:

- Establishing an innovative culture, where experimental activities are commonplace;
- Building and maintaining collaborative and multidisciplinary data science teams;
- Recruiting individual data scientists and using long-term personnel planning;
- Training and supporting personnel in identifying themselves with new competence requirements;
- Aspiring to management and leadership practices, which make these changes possible.

5. There are multiple drivers for exploring the potential of new data sources and technologies.

- First, NSOs need to adapt to the changing digital world and be innovative to stay competitive;
- Second, associated costs of surveys and the increasing non-response rates urge statistical organizations to adapt their production process not to rely on a stovepipe production system;
- Last, but not least, new data and technologies offer an opportunity to do a better job.

6. Statistical information matters for decision-making and fact-checking. Statistics essentially contribute to public debate and allow individuals, households, enterprises and decision-makers to rely on something trustworthy. Therefore, it is important to improve constantly the way in which we produce statistics. Using new data sources and technologies affects the whole production process. At the beginning of the process, statistical

organizations can move towards multi-mode data collection (e.g. combination of phone interviewing, web surveying and sensors) and the use of multiple data sources (e.g. big data and administrative sources).

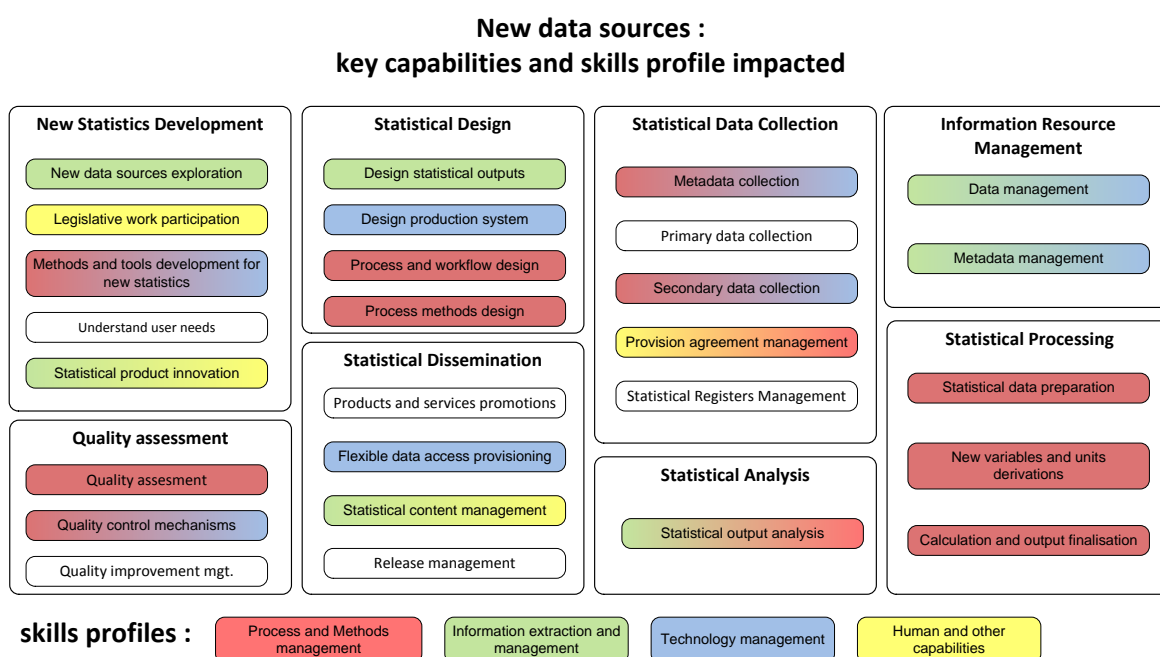
7. This, on the other hand, requires techniques and competences to link, edit and impute data. In the time of new digital data sources, NSOs need the capacity to harness unstructured data of huge volumes efficiently. Examples of new statistical methodologies to be investigated include selectivity correction in big data, pattern recognition, dimensionality reduction and treatment of high-frequency data.

8. The use of new data sources and related technologies will influence a broad range of capabilities of a statistical organization. It does not only require production systems to be updated. Statistical organizations will have the opportunity to provide users with more and more tailor-made data analytics services, including the compilation of sets of indicators tailored to user needs to enable contextualization of information or the provision of data mining tools to allow the user to extract the best value of the statistical data sets.

9. The question is not only *how* to produce statistics but also *what* kind of product to make. Such development in official statistical activities will require the development of an information technology (IT) environment that facilitates the integration of various data sources (data virtualization). The impact of harnessing new data sources across the statistical organization can be visualised using the European Statistical System (ESS) business capability model. The model describes the high-level functions of a statistical organization and allows a systematic and structured approach to the modernization of capabilities in official statistics (see Figure 1).

Figure 1

#### New data sources – key capabilities and skills profile influenced



10. It identifies statistical organization's functions that need to be increased or recreated. Each capability increase will be realized by an appropriate combination of staffing, business processes, methods, information technology (IT) tools and information.

11. To increase these capabilities, NSOs will have to review the staff skills and competences. The modernization requires not only the upgrading of specific profiles, but the 4 groups of profiles, namely methods and statistical processing; information management; IT management; and human capabilities (leadership, creativity and communication).

12. Figure 1 describes the skills profile associated with each of the capability influenced by the use of new data sources. In the past, these profiles used to be well separated and this was usually reflected in the structure of the organization. One of the challenges nowadays is that these profiles will have to be combined synergistically in small teams fitted for agile development. This evolution sheds light on a new type of key function in official statistics, the data scientist or a statistician whose profile can be loosely defined by the intersection of these 4 traditional profiles. In consequence, statistical offices need to consider 1) how to acquire personnel with the needed competence profiles, 2) how modernization affects working methods and organizational culture in statistical offices, and 3) how the management and leadership practices in statistical offices can support the upcoming changes.

13. Next chapter analyses the key requirements related to using new data sources and technologies. After this, the document will discuss the required skills and competences of a future statistician in statistical organizations. Finally, the document will propose ways for acquiring these skills and competences. The document will finish with conclusions and recommendations on the way forward.

## **II. Challenges in harnessing new data sources and technologies**

14. New data sources affect not only data collection phase, but also the entire process of producing official statistics.

15. Gaining access to big data sources and administrative records requires taking into account legal and ethical issues, ensuring confidentiality and establishing a sound process of data delivery with the data providers. In this document, we will consider access to new data sources as a given.

16. In order to be able to utilize new data, one needs to be able to deal with the following methodological challenges:

- Linking data from different sources (surveys, administrative records, registers and big data) and modes (phone and web questionnaires);
- Editing and imputing data with significant volume and velocity;
- Dealing with unstructured data and feature extraction;
- Select the relevant information from huge volumes of data (high-dimensionality, high-frequency or both);
- Deal with computational time issues (feasible algorithms, choice of software);
- Big data selectivity issues (how to ensure representativeness and coverage of big data);
- Pattern recognition;
- Integration of multiple-sources;
- Quality assessment.

17. Most often, addressing those issues with success will require the use of modelling techniques and methods capable of dealing with non-linearity. Among those methods, one can point out artificial neural networks and other machine learning techniques, Bayesian approaches based on simulation and bootstrapping techniques and analysis of network data. Applying such methods implies mastering of optimization algorithms as well as creative and innovative modelling skills. In the same way, integration of multiple sources will also require an increasing use of model-based statistics. Lastly, indicator methods for assessing quality of output statistics will have to be developed in order to maintain the recognized quality standards of official statistics.

18. At the end of the production process, when making new products and tailor-made data analytics the following issues need to be considered:

- Description of multi-dimensional phenomena, which overlap social, economic and environmental domains;
- Integration of qualitative domain knowledge within the analytical framework;
- Contextualisation of statistical information based on discriminatory statistical analysis (identification of relevant sub-sets of data);
- Provision and interpretation of new indicators derived from methods used for producing data;
- Statistical assessment of the sets of indicators;
- Use of new data to promote tailor-made and customized statistics;
- Data visualization and ways to complement traditional dissemination methods with infographics, videos and pictures.

19. Domain expertise and statistical characteristics of given sets of indicators will affect the way in which information can be structured for users. Statistical information needs contextual information and a structure that enables flexibility in the dissemination phase. Furthermore, use of simulation methods or Bayesian approaches can provide, for example, some new types of information, such as complete density functions or state of the observed system. Such information can, thus, offer the users with quantitative scenarios adapted for interpreting and anticipating more complex and uncertain phenomena.

### **III. Competence profile of a future statistician**

20. Competence management plays a crucial role in facing future challenges of official statistics. The chapter will address the following two main questions: 1) what are the competences, skills and features we are looking for in a future statistician, and 2) how can these competences and skills be acquired?

#### **A. Competences and skills for future statisticians**

21. Based on the expected changes in statistical production, described in this document, at least the competences and skills listed in the following table will be needed. The list is by no means exhaustive, but rather points out some examples of such competence areas.

Table 1  
**Competences and skills for future statisticians**

<b>Math &amp; statistics</b> <ul style="list-style-type: none"> <li>• Statistical modelling</li> <li>• Bayesian inference</li> <li>• Machine learning</li> <li>• Identifying and using multiple data sources</li> <li>• Ability to distinguish signal from noise</li> <li>• Capacity to deal with unstructured, massive data</li> <li>• Techniques to link, edit and impute data</li> </ul>	<b>Domain knowledge &amp; soft skills</b> <ul style="list-style-type: none"> <li>• Passionate about the phenomena/substance</li> <li>• Keen on customer needs</li> <li>• Curious about data</li> <li>• Problem solving</li> <li>• Team work</li> <li>• Proactive, collaborative, innovative and creative mind set</li> </ul>
<b>Programming &amp; data base</b> <ul style="list-style-type: none"> <li>• Computer science fundamentals</li> <li>• Scripting language</li> <li>• Languages for statistical computing and graphics</li> <li>• SQL</li> <li>• Map Reduce concepts</li> <li>• Hadoop</li> </ul>	<b>Communication &amp; visualization</b> <ul style="list-style-type: none"> <li>• Collaboration with stakeholders</li> <li>• Negotiation skills</li> <li>• Story telling skills</li> <li>• Visual art design</li> <li>• Visualization tools</li> </ul>

22. Some of the competences listed above have been present in traditional statistical work for years. Thus, data science has rather an effect on the weight of the different competence areas in a future statistician's competence profile.

23. The availability of new digital mass datasets changes the traditional paradigm of statistical production. In the near future, statistical production more often starts by data scientists digging and analysing unstructured datasets and thinking of potential ways to use them. This is the opposite of the traditional approach, where one starts by specifying the needs and then moves to data collection. For statisticians, this requires an ability to adopt a flexible and proactive data driven approach to discover the informational content of new data sources. In this context, using the concept "a hacker's mind set" describes well how statistical work will require re-evaluation of problems and continuous exploration of the limits of what is possible.

24. Another example of how emphasis on competences has changed is story telling. Future statisticians need to be able to present their figures in a context and communicate statistical information visually. The emphasis is, hence, slightly different from producing numbers per se and providing researchers with data for their further analysis. A future statistician needs to be able to produce value added by providing analysis, insights and context to the figures produced. In support of this, skills in visual art design and using different visualization tools will become more significant as well.

25. Statisticians in statistical organizations most often possess the ability to learn and apply data science techniques in an experimental way. Nevertheless, new challenges in statistical production will also require the following skills:

- The ability to combine information from different sources in order to produce new information products. This requires substantive knowledge of the phenomenon to be described and a good understanding of the customers' needs;
- Effective and efficient communication with other stakeholders for building and testing new methods, in particular for collaboration with different domain experts and academic institutions;
- Programming skills to build robust and scalable data processing and analysis tools and to use scriptable statistical languages that enable application of new methods.

## **B. Means for acquiring skilled statisticians and supporting data science-driven work**

26. Data science requires many skills that are seldom found in one employee. Thus, the mastery of data science begins to flourish in collaborative and multidisciplinary data science teams that consist of experts with different competence profiles and backgrounds. Efficient data science teams require the possibility to work horizontally which requires a certain level of flexibility from the organization and its leadership. However, the duties and responsibilities of teams have to be jointly agreed and understood. This applies also to management practices: the duties and roles between directors, supervisors and team leaders have to be defined and applied. Thus, increasing teamwork in statistical offices challenges organizations to find ways of building and maintaining teams that cut across the organization, utilizing best practices and learning from each other.

27. Although data science teams consisting of various experts will probably be one of the most typical ways of organizing data science-driven work, it might be necessary to recruit individual data scientists to work, for example, with the following tasks:

- Conducting plans aiming to increase the use of new digital mass data in statistical production.
- Identifying new data sources and evaluating their potential.
- Developing processes, methods and IT solutions.
- Training and consulting, sharing competence in data science issues.

28. Considering recruitment of data scientists, like all recruitment activities, the brand of the statistical organization is important and needs to be appealing to the future employees. Even more importantly, whether recruiting individual data scientists or experts to the data science teams, organizations need to have a realistic understanding of tasks and competence profiles of such staff. Long-term personnel planning helps to define in what extent data science experts are needed and what kind of positions will be available. There will be a great demand for data scientists and retaining them will not be easy. Therefore, one must ensure that the data scientist's tasks correspond to his/her competences. It may be challenging to retain data scientists in any case, especially if they are recruited for positions where they will not be able to use their full potential.

29. One way of approaching data scientists, partially connected to recruitment, is collaboration with universities. Statistical offices can recruit trainees or thesis workers with a computer science background. Providing students with real datasets and practical job opportunities, offers an opportunity for statistical offices to promote themselves as an interesting place to work. In addition, collaboration with universities helps understand the potential of data science and the skills that are relevant in tackling the challenges of statistical production.

30. When considering different alternatives for acquiring competence from outside of the organization, recruitment is not the only option. Statistical offices do not necessarily need to own the competences and skills themselves. They can also acquire the competences by relying on their networks or outsourcing some tasks. Therefore, the statistical community will have to create efficient partnerships with research and data science networks, as well as with the relevant experts and users of statistics. Collaborations would ideally offer new opportunities for statistical organizations to keep their competences up-to-date.

31. One of the most traditional ways of competence development is training. Once future competence needs and current gaps have been specified in the organizations, they

can organize themselves or acquire training courses, workshops and programmes to address the specific competence gaps. However, training is not sufficient as the only tool of acquiring competence; it is rather a complementary tool. It is also important to bear in mind that the current personnel might already have an existing interest and potential in data science issues. This potential may only be waiting to come out when given a chance to get familiar with new datasets and try new methods. In order to support staff's abilities, organizations should experiment with new datasets and allow some time for doing so. Once again, like in team building, decisions, prioritization and support from directors and supervisors are needed.

32. Very often, the needed changes, such as building teams across the organization or creating necessary management or work environments, are in contradiction with current stovepipe approaches in statistical offices. Therefore, realizing these changes call for the ability and willingness to do things differently, think in a creative and innovative way and take courageous steps forward. For example, using new datasets requires a good degree of imagination; one often cannot specify the usefulness of the dataset before analysing it, and one needs an environment that allows creativity and development of new information products. Statistical organizations have to aim at an innovative culture where experimental activities are commonplace and permission to fail is an essential part of the culture.

33. An experimental culture challenges the present, the way in which statistics have been produced for a long time. Statisticians' professional identity relies on producing objective and reliable statistics of good quality. Having a tradition is significant for statisticians that have been accustomed to work as "the memory of a nation". This remains a valuable task, even when the challenges from our operational environment demand statisticians for a change. While taking care of the production of high-quality statistics, we also need to take care of professional statisticians. As statisticians, we also need to be encouraged and supported in redefining ourselves with new competences.

34. The statistical community is already considering new ways and methods to produce statistics. This has to be linked with what the needed competences are and how these can be acquired. The fact is that we are all official statisticians, whose primary unchanging mission is to make good statistics better. In addition, the question is not how a statistician can become a data scientist but how a statistical organization can include "data sciences" in its toolbox and how the needed competences and skills can be fused into the professional identity of official statisticians.

## **IV. Conclusions and recommendations**

35. The document identifies several ideas for the future with a view to adapt statistical organizations to the digital era. The following are the key points and recommendations:

- Harnessing new data sources and developing new data analytics in official statistics is a key objective for remaining competitive and making good statistics better;
- New methods and changes in the paradigm of producing statistics affect the staff competence requirements; making the development of skills in data science crucial;
- Staff of statistical offices have to be encouraged and supported in redefining themselves with new competence requirements and new working methods;
- Needed competences and skills can be acquired for example by training the current staff, recruiting individual data scientists and collaborating with universities and other networks;



- Upgrading data science skills is necessary, but as important is changing the working methods and organizational culture of statistical offices;
- Future working methods have to be more collaborative, horizontal and multidisciplinary; data science teams that cut across the organization to include the necessary skills will be the way to organize data science-driven work;
- Statistical offices need to adopt an innovative culture where experimental activities are commonplace; this is important since future statistical work requires creative thinking, as well as an ability and will to do things differently;
- Management and leadership have a crucial role in realizing upcoming changes, which require, for example, long-term personnel planning, building and maintaining team work, allowing time and space to carry out experimental activities, and using participatory ways in defining the common future in statistical offices.

## V. References

C. F. Jeff Wu: *Statistics=Data Science?* University of Michigan, Ann Arbor. [www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf](http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf);

*Methodology architecture: A roadmap for new methodological directions in the Australian bureau of statistics.* Statistical Journal of the IAOS, vol. 30, no. 4, pp. 371-375, 2014

*ONS Strategy 2013–2023, Trusted Statistics – Understanding the UK.*

<http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/about-ons/what-we-do/corporate-documentation/strategies-and-policies/ons-strategy-2013-23/index.html>

---