UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

ECE/CES/BUR/2015/FEB/11
20 January 2015

CONFERENCE OF EUROPEAN STATISTICIANS

For discussion and recommendations

Meeting of the 2014/2015 Bureau
Geneva (Switzerland), 17-18 February 2015

Item 3 (e) of the Provisional Agenda

OUTCOMES OF THE PROJECTS BY THE HIGH-LEVEL GROUP FOR THE MODERNISATON OF STATISTICAL PRODUCTION AND SERVICES IN 2014

Note by the High-Level Group

The note provides a summary of the results of the projects overseen by the High-level Group for the Modernisation of Statistical Production and Services during 2014. ***The Bureau reviewed the work done and provided advice on future work.***

I. INTRODUCTION

1. The High-level Group for the Modernisation of Statistical Production and Services (HLG) was created in 2010 by the Bureau of the Conference of European Statisticians. It comprises the heads of ten national and international statistical organizations, and has a mandate to reflect on and guide strategic developments in the ways in which official statistics are produced.

2. Each year HLG organises a Workshop, inviting representatives of various groups and projects related to modernisation of official statistics. These workshops help to ensure coordination of activities. They also review progress and determine the key priorities for the following year. In 2013, the HLG workshop decided that **implementing the Common Statistical Production Architecture and investigating the potential use of Big Data sources for official statistics were the two highest priorities in 2014**. The HLG launched international collaboration projects to address these priorities, which ran during the calendar year 2014. This paper summarises the main results of these projects.

II. COMMON STATISTICAL PRODUCTION ARCHITECTURE (CSPA)

3. **The Common Statistical Production Architecture (CSPA) was created during 2013 in an international collaboration project under HLG.** That project included the development and specification of the architecture, as well as a practical test of its principles and applicability in a “proof of concept”. However, given the strict time limits for that project (one year), it deliberately did not include the implementation of the architecture for production systems within statistical organisations.

4. The CSPA provides a standard framework for developing the components of statistical production, in a way that they can be much more easily shared within and between

organisations than has previously been the case. It builds on key international standards such as the Generic Statistical business Process Model (GSBPM), and the Generic Statistical Information Model (GSIM).

5. The CSPA implementation project ran during 2014, and had the following aims:

- Implement the CSPA in practice by creating CSPA-compliant services that could be shared between processes and organisations;
- Develop the resources necessary to support CSPA implementation, including training materials, and the proposed catalogue of services and other artefacts;
- Further test the applicability of the GSIM, and, if necessary, to suggest further refinements to that model for a possible future revision.

6. During the project the following CSPA-compliant services were developed:

- Seasonal adjustment – France, Australia, New Zealand
- Confidentialized analysis of microdata – Canada, Australia
- Statistical chart generator – OECD
- SDMX transformer – OECD
- Sample selection – Netherlands
- Linear error localisation – Netherlands
- Linear rule checking – Netherlands
- Error correction – Italy

7. Two additional services were specified, ready for building during 2015

- List statistical classifications – Norway
- Retrieve statistical classifications - Norway

8. The first eight services have been incorporated into a “service catalogue”, which was specified by the CSPA Catalogue Task Team. It was developed as a prototype for demonstration purposes by UNECE, and then implemented as a “live” version by Eurostat. Several of these services are already in use in statistical production, e.g. the seasonal adjustment service has been used by the Netherlands.

9. The specification and development of these services was overseen by an “Architecture Working Group”, chaired by New Zealand. This group has advised the service developers whenever they had a question about the interpretation and use of the CSPA.

10. A tangible result of the work of the Architecture Working Group, is a slightly revised version of the CSPA (version 1.1), which was produced within the project, and has been released for users[1]. It incorporates additional guidance based on the lessons learned from the project.

[1] http://www1.unece.org/stat/platform/display/CSPA

III. BIG DATA

11. This project had three main objectives:

- To identify, examine and provide guidance for statistical organizations on the main possibilities offered by Big Data and to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry
- To demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts
- To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

12. To deliver the objectives above, the project established the following work packages:

- Work Package 1: Issues and Methodology
- Work Package 2: Shared computing environment ('sandbox')
- Work Package 3: Training and dissemination
- Work Package 4: Project management and coordination

13. Three ‘task teams’ were established under Work Package 1, to address the Big Data issues identified by participating organisations as the most important for official statistics: Privacy, Partnerships and Quality. Practical results in Work Package 2 were delivered by the Sandbox Core team and eight ‘experiment teams’, working with different types of Big Data. Several training events were held under Work Package 3, mainly introducing statistical organisation staff to Big Data software tools. The training materials are available for re-use. A survey of statistical organisations identified the skills needed to work with big Data and the areas where training was most needed to develop those skills.

A. Summary of Findings

14. The findings for each part of the project are presented below. All outputs are available at http://www1.unece.org/stat/platform/display/bigdata/2014+Project.

Quality

15. The Quality Task Team assessed existing quality frameworks for official statistics with respect to their applicability to Big Data (including the quality frameworks of Statistics Sweden, Statistics Canada, the Statistical Network, the Australian Bureau of Statistics, and the European Statistical System Code of Practice). They found that:

- there is a need for quality assessment covering the entire business process;
- input quality can be explored and assessed by using and elaborating existing input quality frameworks;
- throughput quality can be maintained by following quality processing principles, but throughput quality dimensions need to be further developed for Big Data processing;
- additions have been proposed to output quality dimensions from existing frameworks, to make them suitable for Big Data applications: This needs further testing.

16. The main output of the Quality Task Team was a framework for the quality of Big Data.

Privacy

17. The Privacy Task Team was asked to give an overview of existing tools for risk management in view of privacy issues, to describe how risk of identification relates to Big Data characteristics, and to draft recommendations for statistical organizations on the management of privacy risks related to the use of Big Data. They found that:

- Existing tools are well-developed;
- Privacy risks can be linked to Big Data characteristics;
- Recommendations have been formulated on information integration and governance, statistical disclosure limitation/control and managing risk to reputation;
- But: **There is not much experience yet with Big Data privacy issues**.

18. The team produced papers taking stock of the current status of statistical disclosure control, investigating the characteristics of Big Data and their implications for data privacy, and a summary of practical measures to manage Big Data privacy.

Partnerships

19. A Task Team was set up to examine the issues around partnering with different types of organizations within a Big Data context. Two questionnaires were sent to national and international statistical organizations. The first focused on the overall strategy for Big Data in the organization, while the second focused on specific projects on Big Data. The task team also received information from the Sandbox experiment teams about their practical experience with data providers, technology and academia partners. They found that:

- Most partnership arrangements encounter similar forms of issues related to financial/contractual arrangements, legislative, privacy and confidentiality issues, responsibilities and ownership issues and other risks;
- The importance of these issues depends on the type of partner;
- For the Sandbox experiments, the main issue was timely access to data;
- A project can only exist if a working partnership can be forged with a data provider to provide a reliable data source.

20. The main output of this task team was a set of guidelines for the establishment and use of partnerships in Big Data projects for official statistics

Sandbox

21. A web-accessible environment for the storage and analysis of large-scale datasets was created with support from the Central Statistics Office of Ireland and the Irish Centre for High-End Computing (ICHEC), and used as a platform for collaboration across participating organizations. This environment, known as the “sandbox” provides a computing environment to load Big Data sets and tools, with the goal of exploring how they could be used for statistical production. The Sandbox Task Team found that:

- A common computing environment enables shared work on methodology, especially where the data sets have the same form in all countries. Methods can be developed and tested in the shared environment and then applied to real data sets within each organization. For example, a methodology for sentiment analysis developed in the Netherlands was tested on data from Mexico, and a program for treating smart meter data from the United Kingdom was applied to Canadian data with relative ease;
- Although web sources and social media are appropriate for international sharing, language differences can present a problem when cleaning and classifying text data;
- Other work on methodology was done in the mobile phones team, for computing the movement of people starting from call traces, and in the traffic loops team, for calculating the average number of vehicles in transit for each day and for each road;
- The project shows for the first time, on a practical basis and on a broad scale, the potential and the limits of the use of Big Data sources for computing statistics. Improvements in efficiency and quality are possible by replacing current data sources. New products can be obtained from novel sources such as traffic loops, mobile phones and social media data. However, some sources can be of low quality and require some serious pre-processing before use. In general, Big Data sources can be effectively used as additional sources, benchmarks or proxies;
- The same Big Data sets can be used in several contexts and for different purposes;
- The use of a shared environment for the production of statistics is severely limited by privacy constraints on data sets. These constraints often limit the personnel authorized to access and treat the data, and do not allow files to be moved outside the physical boundaries of a single organization;
- These limitations can be partly bypassed through the use of synthetic data sets. Another solution is to generate the data by modelling its behaviour. Both approaches were used in the project, for smart meters data and scanner data respectively.
- Big Data tools are essential when the size of data sets is measured in terms of hundreds of gigabytes or larger. They can be more efficient than existing data processing tools for data larger than one gigabyte;
- Researchers/technicians should be able to master different tools and be ready to deal with immature software, so strong IT skills are needed.

22. The main outputs of the Sandbox Task Team were a series of experiment reports giving detailed information about the conduct and conclusions of the different experiments conducted.

* * * * *