Using New Data Sources to Modernise Official Statistics – A Collaborative Approach

Jenine Borowik*
Australian Bureau of Statistics, Canberra, Australia – jenine.borowik@abs.gov.au

Thérèse Lalor
Australian Bureau of Statistics, Canberra, Australia – therese.lalor@abs.gov.au

Peta Darby
Australian Bureau of Statistics, Canberra, Australia – peta.darby@abs.gov.au

1. Introduction

Many statistical organizations are struggling to find the resources to continue to produce robust and relevant statistics to meet the increasing demands of customers. There is a growing need for organizations to respond to a number of challenges in the current statistical environment in order to deliver efficiencies, reduce provider burden and make richer use of existing information sources. As the statistical environment changes, technological developments are also expanding the possible tools for data collection.

Some of these challenges are directly relevant to data collection, and include:
- the increasing costs of traditional Censuses and surveys, a common model for statistical production in statistical organizations;
- quickly responding to emerging information needs;
- effectively using administrative data and harnessing alternative data sources;
- efficient processing, integrating into a multi-source environment, privacy and security issues, and cooperation and partnerships with those outside official statistics;
- developing innovative and automated approaches in order to integrate the new technologies into data collection processes.

The statistical environment is changing at a global level, and challenges all statistical organizations in various ways. Organizations need to adapt to the world of the future – where data are available in abundance from many sources, and sometimes on a close to 'real time' basis. To successfully and effectively harness the challenges and opportunities presented in this new environment, the official statistics community will need to draw on its strengths through international collaboration. Failing to address these challenges and opportunities threatens the relevance of official statistics and the vital role it plays in informing decision making around the world.

2. Responding as a community

We are all facing modernisation challenges though the nature of these challenges may vary from one country to another. In that sense modernisation is a relative term. As a community – there is a lot we can learn from each other. We need to work together to develop cost effective methods of modernising. We can learn from the experience and mistakes of others thus making the path towards modernisation easier for those that will follow when they are in a position to do so.

The High Level Group (HLG) for the Modernisation of Statistical Products and Services, which consists of 10 heads of national and international statistical organizations, was established in 2010 to drive and foster collaboration between statistical organizations. This group recognises that statistical organizations needed to re-invent our products and processes and adapt to a changed and changing world.

3.   The common languages for collaboration

HLG has focused to date on creating the frameworks and standards that have enabled and will continue to enable us collaborate into the future. The most notable of these standards are the Generic Statistical Business Process Model (GSBPM), the Generic Statistical Information Model (GSIM), and the Common Statistical Production Architecture (CSPA). An illustration of the relationship of these three standards to the modernisation of official statistics is shown in Figure 1 below.
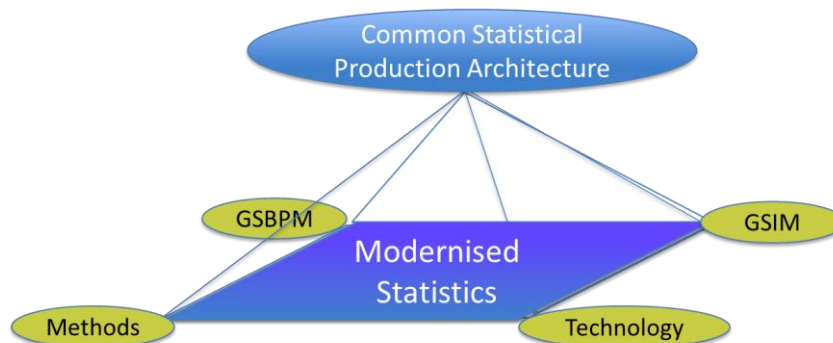

Figure 1 Relationship between HLG standards and modernised statistics

The HLG identified the GSIM as a key standard, in partnership with the GSBPM, to drive the modernization of official statistics. The GSBPM describes the process of producing statistics and GSIM provides a common language to describe the input and outputs of the processes.

The CSPA builds on the GSIM and GSBPM, and provides a common reference architecture that enables statistical organizations to develop and standardize components of statistical production, regardless of their origin. This allows the components to be much more easily shared and reused both within and between organizations than previously possible.

4.   Achieving results through collaboration

GSBPM, GSIM and CSPA provide a common conceptual basis for collaboration work. This helps to overcome one of the biggest challenges for collaborative work. A common way of referring to statistical production helps ease the communication between organizations.

All resources in HLG work are voluntary. This is typical of most collaborative work. People are often contributing to the work in addition to their regular work. With the wide range of time zones involves, participants are often attending meetings out of office hours. To maintain enthusiasm, the priority areas reflect as far as possible the needs of the participating organisations. This can mean that there is more delivered in certain areas where there was a strong interest from participants, and slightly less in a few other areas. This flexibility to respond to evolving requirements and changing priorities is a key factor in the success of the projects.

Governance of collaboration can also be a challenge. The HLG has adopted an appropriate governance structure to ensure efficient and effective implementation of its strategy. This structure includes the Executive Board, to which a number of projects such as the CSPA and Big Data projects report, as well as four Modernisation Committees, for Organizational Framework and Evaluation, Production and Methods, Products and Sources, and Standards. An illustration of the structure is shown in Figure 2 below.
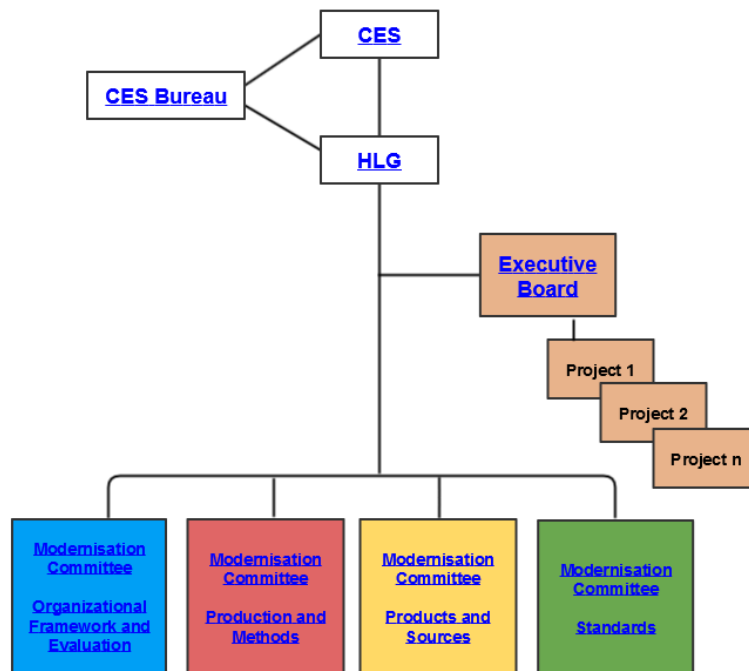
Figure 2 High-Level Group Governance

As part of their current program, the HLG sponsors two key strategic collaboration initiatives that respond to the emergence of new data sources. These include a Modernisation Committee on Products and Sources, which are tasked with examining and exploring new data sources at a closer level, and a project on Big Data in Official Statistics, which aims to identify possible and feasible opportunities for official statistics in Big Data.

5.   Modernisation Committee on Products and Sources

The Modernisation Committee on Products and Sources considers and makes proposals on how to develop the range of sources needed to support modernised statistical production and services, as well as the various products needed to meet the increasing demands of users. It has operational responsibility for work on statistical data collection, including Big Data, and various aspects of dissemination including access to micro-data, statistical confidentiality and open data.

Some of the key activities of this group related to data collection are discussed below:

Big Data Inventory

Through a series of task teams, a significant amount of information about the use of Big Data in various national and international statistical organizations has been gathered. This has been published as a useful resource for the official statistics community. The inventory can be accessed by visiting www1.unece.org/stat/platform/display/bigdata/Big+Data+Inventory.

Mobile Devices
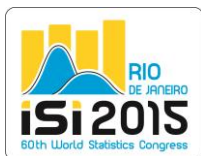
The expansion and increased use of mobile devices (smartphones, tablets, wearables) and the progress made in technology innovation provides new approaches and opportunities for increasing the frequency and efficiency of data collection and dissemination of official statistics.

The use of mobile devices as an instrument for data collection enables the completion of traditional surveys at an increased speed and improved quality. For example, the Global Positioning System (GPS) information transmitted from mobile phones can be used to produce better and more timely time usage, travel and transport statistics. There are many opportunities to be explored and harnessed with this emerging data source.

The Modernisation Committee will undertake work to analyse some of the base case studies and experiences on the use of mobile devices to establish some guiding principles as well as technical standards.

Administrative Data
The Modernisation Committee has taken on a task team exploring methodologies for an integrated use of administrative data in the statistical process. To date, this team have produced:

- a mapping of how administrative data can be used in the statistical production process using GSBPM as the foundation; and
- a framework to assess the quality of administrative data and its statistical usability – prior to statistical usage, to determine if an administrative data source can be used for statistical purpose and how.

Other areas of focus
The Modernisation Committee is also focussing its efforts on other areas, including:
- Linked Open Data, with the aim of formalizing and publishing some of the metadata used in official statistics as open data. This presents some significant challenges, but when achieved will make the metadata visible, globally identified, linkable and more easily actionable;
- Mixed Mode, which can present operational and quality challenges; and
- Communication, including communication with providers, and communicating the value of official statistics.

The Modernisation Committee on Products and Sources will continue to explore these opportunities throughout 2015.

6. Project on Big Data for Official Statistics

The HLG strategy document[1] states that "products and services must become easier to produce, less resource-intensive, and less burdensome on data suppliers" and that "new and existing products and services should make use of the vast amounts of data becoming available, to provide better measurements of new aspects of society".

---

[1] Strategy to Implement the Vision of the HLG available at
www1.unece.org/stat/platform/display/hlgbas/HLG+Strategy

In 2014, the HLG sponsored a project aligned with these aspirations with a focus on the role of Big Data in Official Statistics. The project involved 75 participants from 20 different organizations, and had the following objectives:

- to identify the main possibilities offered by Big Data to statistical organizations; and
- to demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources.

The project involved four task teams each focussed on a different facet of Big Data in Official Statistics.

The Quality Task Team investigated the implications of Big Data for the quality of official statistics, and developed a preliminary quality framework for Big Data.

The Privacy Task Team were tasked with providing 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.

The Partnership Task Team was set up to examine the issues around partnering with different types of organizations within a Big Data context. Partnership with data providers and sources is an important and often first step, but in order to optimise the use of the data, other types of partnerships might be needed. Statistical organizations might need to collaborate with each other but also to partner with organizations such as academia, scientific communities, research institutes, and technology providers, not only to develop standards, processes and methodologies, but also to gain access to organizations with analytical capability and access to the most advanced technology. General guidelines were prepared by the team on establishing partnerships for Big Data.

The work of the Partnerships task team was underpinned by the information from a questionnaire developed jointly with the United Nations Statistical Division, in preparation for the International Conference on the Use of Big Data for Official Statistics, which was held in Beijing, China, in October 2014

The project also had a significant practical element - hands-on work on a Big Data tools and dataset in common, shared computation environment. The objectives of the sandbox were to:

- Explore tools and methods;
- Test feasibility of Big Data derived statistics; and
- Replicate outputs across countries.

The sandbox was created with support from Central Statistics Office (CSO) Ireland and the Irish Centre for High-End Computing (ICHEC). In the Sandbox there was an excellent cooperation with technology providers for example ICHEC and Hortonworks. Data was provided or facilitated by several national statistical organizations, including Canada, Ireland, Italy, Mexico, and the Netherlands, as well as by the mobile provider Orange.

The HLG is sponsoring an extension of the Big Data project in 2015. The main focus is the work in the sandbox. This continues the experiments started in 2014 and aims to produce a multi-national statistic from Big Data sources.

7. Conclusion

Effective collaboration in sharing knowledge, expertise, tools and methods is vital for the official statistics industry to modernise effectively and as quickly as possible. There has been strong progress made already in collaborations related to data collection, with a number of tangible results. These achievements would not have been possible without the excellent cooperation with many organizations and companies.

Moving forward, it is important to continue to working together to improve the efficiency of the statistical production process and the ability to produce outputs that better meet user needs.