Big Data Development at the Global Level
The UN Global Working Group on Big Data

Niels Ploug[i]

Statistics Denmark & Co-chair of the Global Working Group

Abstract. This paper introduces the work of the UN Global Working Group on Big Data for Official Statistics. This work is organized around a number of Task Teams and International Conferences. At the 2017 conference in Colombia the Bogota declaration on 'trusted data partnership' was formulated and the Global Working Group started its work on The Global Platform for trusted data for official statistics.

1. Introduction

The UN Global Working Group (GWG) on Big Data for official statistics was created in 2014 by the UN Statistical Commission to explore the benefits and challenges of the use of new data sources and technologies for official statistics and SDG indicators. The purpose of the GWG is to addresses issues pertaining to methodology, quality, technology, data access, legislation, privacy, management and finance. The GWG consists currently of 28 member countries and 16 international organizations.

The GWG's activities are led by a bureau and organized in Task Teams many of which has a focus on a specific data source. On top of this, the GWG has organized a yearly International Conference on Big Data for Official Statistics since 2014.

One of the first activities of the GWG was to create a big data inventory with input from NSO's worldwide. The inventory contains a short description of big data projects and other activities that take place at NSO's and can be used for inspiration by NSO's in other countries. The inventory is available at the GWG's webpage: https://unstats.un.org/bigdata/inventory/

This paper give an introduction to the work of the GWG Task Teams, an overview of the content of the International Conferences and a description of the latest developments in the GWG's work to establish a global platform for big data and the GWG's collaboration with tech companies.

2. Task Teams

The Task Team members are NSO's, international organizations and tech companies. A Task Team chair leads the work. The Task Teams produce guidelines for the use of specific big data – data sources in the production of official statistics based on the experiences of the Task Team participants. Task Teams organize training sessions in relation to the GWG's International Conferences and present their work at the conferences for inspiration and discussion.

This section give a short presentation of the work of the GWG's Task Teams.

Access and Partnerships

The purpose of the work of the Task Team on access and partnership is to explore the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working.

The objectives of the Task Team are to facilitate access to Big Data sources for official statistics and facilitate forming partnerships with other public and private organizations in order to work with Big Data. It is important that this is done in a way that reflects a mutual understanding with partners of what is reasonable to expect from each other, by respecting each other’s position, role, aims, business model, social responsibilities, limitations and possibilities.

Access to Big Data sources and forging partnerships with other public and private organizations in order to work with Big Data is becoming ever more important to national statistical systems (NSS) for fulfilling their mission in society. The national statistical systems (NSS) has an interest to collaborate with the private sector, in order to advance the potential of official statistics. One important purpose of the work of this Task Team is to communicate the advantages of exploiting the wealth of available digital data to the benefit of society by using them as input in the production of trusted statistics with equal access for all.

Mobile Phone Data

Mobile Phone Data has surfaced in recent years as one of the Big Data sources with a lot of promise. It is expected that Mobile Phone data could fill data gaps especially for developing countries given their high penetration rates. In its 2014 Measuring the Information Society Report, ITU showed that the average mobile subscription rate is 96.4 per 100 inhabitants worldwide, with some lower averages in Asia (89.2) and Africa (69.3). Nevertheless, these numbers show how pervasive mobile phone use is. ITU elaborates that rural areas are still lacking behind urban areas, and this should be considered in studies using Mobile Phone data, but it is clear that the coverage of these data is global. Almost every person in the world lives within reach of a mobile-cellular signal.

The purpose of the work of the Task Team is to execute one or more pilot projects on the use of mobile phone data for official statistics and to document and publish the lessons learned, as well as prepare and maintain knowledge base on the use of mobile phone data.

Satellite Imagery and Geo-Spatial Data

The Task Team on Satellite Imagery and Geo-Spatial Data aims to provide strategic vision, direction and development of a global work plan on utilizing satellite imagery and geo-spatial data for official statistics and indicators for post-2015 development goals. It will build on precedents to innovatively solve the many challenges facing the use of satellite imagery and geo-spatial data sources.

Statistical agencies around the world have a strong interest in investigating the viability of using satellite imagery data to improve official statistics on a wide range of topics spanning agriculture, the environment, business activity and transport. Satellite imagery has significant potential to provide more timely statistical outputs, to reduce the frequency of surveys, to reduce respondent burden and other costs and to provide data at a more disaggregated level for informed decision-making.

The Task Team has drafted a comprehensive guideline for the use of satellite data in the production of official statistics that after an editorial process will be made available for NSO’s and other potential users of satellite data.

Scanner Data

Scanner data is a Big Data source being increasing used in national statistical systems for the calculation of price indices as statistical offices explore ways to meet the expectation of society for enhanced products and improved, more efficient ways of working. Many of the price measurement issues and methods for scanner data from supermarket chains and other retailers apply also to other big data sources (for example, online prices obtained from web scraping).

The Task Team work on the following deliverables

1. The delivery of an open source application with an associated Application Program Interface (API), which can be shared among all partners in the statistical community. This application will take cleaned and classified scanner data (i.e. Analysis Ready Data (ARD)) and will apply a range of analysis and monitoring processes before enabling a range of methods to be used for estimation of price indexes. The user can specify the exact method.

2. The development of training and instructional material on the use of the application

3.The development of accompanying methodological guidance material which will a) summarize the relevant literature on methods, b) point to internationally-agreed recommendations on which methods are appropriate in which situations and c) catalogue existing and intended practice across NSIs in the use of Prices big data.

Social Media Data

In its 2014 measuring the Information Society Report ITU estimates that 40% (almost 3 billion people) were using the Internet by the end of 2014, meaning that 60% were not using the Internet. Moreover, Internet usage varies considerably across regions, ranging from 75% in Europe to only 19% in Africa.

Against this background, the task team work to clarify which kind of social media data can be collected, how it can be collected, how it can be analyzed and processed into statistics, useful for policy purposes.

Training, Skills and Capacity-building

Big Data is by definition different from traditional data sources used by national statistical systems (NSSs). This implies that new methodologies need to be developed to work with Big Data. The kind of sources of Big Data poses challenges both in how to approach their processing and analysis, but also the mere technological way of dealing with them. This means that new skill sets are necessary to successfully work with the new Big Data sources. Some of these new skill sets could be hired temporarily, others will need to become in an integral part of the institution.

An additional complication is that there is not just one kind of Big Data source, and each kind of Big Data may have different requirements as far as new skill sets are concerned. It is therefore a separate task to develop tools to identify and assess the needs for new skills.

The objectives of the work of the Task Team is to develop methods and tools (including coordination and facilitation) for baseline need identification of Big Data skills in NSSs; and to perform an assessment of institutional readiness of NSSs in using Big Data. On top of this to provide guidance on the development of a program of training that addresses the gaps identified in the skills needs analysis - and to facilitate establishing global network of institutions for training and capacity building on Big Data.

3. Global Conferences and Open Day of The Global Working Group

The GWG organized its first international conference on Big Data for official statistics in China in 2014 followed by the 2015 conference in Abu Dhabi, 2016 in Ireland and the latest 2017 conference in Colombia.

The theme of the conference in Colombia was ‘data collaboratives and trusted data’, which in short sets the scene for the direction of the work of the GWG since the conference.

Data collaboratives are a new challenge and new opportunity for the community of official statistics – in relation to Big Data, to the SDGs, to the sharing of data, services, technologies, and know how.

The following crucial questions where discussed during the conference through among other things a number of interventions that showed good examples for data collaboration and the use of Big Data from NSO’s as well as major tech companies:

•How can statistical offices, technology companies and data owners collaborate in a mutually beneficial way in a changing world, in which data are seen as the most important source for creating wealth and development for all?

•What are the experiences and lessons learned from existing data collaboratives in relation to coverage, inclusion (and exclusion) of partners, activities, management and financing?

•How can we share micro-data and other sensitive data in a federated cloud environment given regulatory frameworks for data privacy and statistical laws protection confidentiality?

•How could we effectively and collaboratively use modern tools and services, such as data lakes, integrated geo-spatial data and statistics, or open source elastic stack while adapting job profiles and skills sets in the statistical office?

As part of the conference was formulated the Bogota Declaration[1] addressing the future work of the GWG.

In the declaration it was proposed to : *provide a major thrust for the strategic area of the Cape Town Action Plan on innovation and modernization by advancing global data collaboratives, facilitated by a trusted federated global platform initially for research and development in the discovery, access and use of data, statistical methodology, software applications and capacity building for the production of statistics and indicators. These partnerships will innovate and help modernize official statistics and their use of new data sources, including Big Data. It will enable data driven transformation in the production of specific statistics or SDG indicators for better decision making.*

The development of the global platform has been one of the most important parts of the work of the GWG since then.

The GWG has organized an ‘open day on the global working group’ on the 21st of October in Dubai (the day before the opening of the 2nd World Data Forum) where the Global Platform will be presented.

4. The Global Platform and Collaboration with Tech Companies

The purpose of the Global Platform is to take the work of the GWG to a new level.

[1] https://unstats.un.org/unsd/bigdata/conferences/2017/Bogota%20declaration%20-%20Final%20version.pdf

The GWG started its work by mapping the experiences of NSO’s work with Big Data. This knowledge is available in the inventory mentioned in the beginning of this paper. Based on this the Task Teams was created.

Based on the experience that far the GWG decided that there is a need to facilitate the exchange of experiences of the use of big data sources, projects, knowledge and algorithms.

The GWG wants to prove over the next 18 months that data collaboration on the UN Global Platform with involvement of large tech companies is possible and is to the benefit of all statistical office. To achieve this the GWG has started up several projects. With Google it engaged on estimating environment statistics and (SDG) indicators use satellite data and Google's Earth Engine. GWG agreed with Nielsen using their price data for various products from around the world to calculate price fluctuations. Recently, GWG agreed with the "Tech for Social Impact" arm of Microsoft to start up possibly four projects: on estimating agricultural crop and yield statistics in Canada and Poland; on estimating land cover and land cover change together with FAO; and two projects on measuring human mobility (especially for migration statistics) with Georgia and Indonesia respectively. These collaborations include many players from the statistical community, but also from academia, research and private sector. The expected outcomes of these projects are common methods, algorithms, tools and APIs, especially also for the estimating of SDG indicators, plus lessons learned on data sources, partnership arrangements, business models and confidentiality and data security issues.

The Global Platform should provide scale and scalability for access to and capacity building in the use of Big Data and its integration with administrative sources, geospatial information and traditional survey and census data, and for use of the related services, applications and infrastructure.

In terms of products and services, the Global Platform could:

(a) facilitate the access to and the use of information technology infrastructure and proprietary data sources;

(b) host specific global projects such as on the global enterprise registers and on bilateral or multi-lateral data asymmetries;

(c) store and give access to new open source applications for the use of multi-source data sets for more timely and flexible production processes for official statistics and

(d) provide training and capacity building services in the use of new data sources in the statistical production. Being a global platform for official statistics, it envisaged that the data access could also include access to trusted (micro) datasets from official statistical sources and access to verified and curated data provided by private and public partner organizations. Services of the global platform could include services of Cloud storage and computing, technical statistical services (e.g. stand alone or as building blocks to develop applications), and capacity building in relation to these services.

As an example of how the provision of services and applications could work, a project could be initiated to transfer the methodology and provide access to the data and technology infrastructure of using satellite imagery data for the calculation of agricultural statistics in either a developed or developing country environment. Such a project requires

(a) access to satellite data for training and testing purposes;

(b) preparation of the data in a secure and appropriate IT environment;

(c) testing of the methodology;

(d) compilation of the agricultural statistics; and

(e) training of the staff to be able to do so on a routine basis.

The Global Platform as a network of partners (statistical offices, research institutes and private companies) and through its technology infrastructure would be able to offer and deliver these services, including the services for training and capacity building.

In conclusion, the GWG works to establish a Global Platform for data, services and applications under auspices of the Commission. This platform could be a network of partners (statistical offices, research institutes and private companies) with appropriate technology infrastructure which would be able to offer and deliver data, applications and services for the community of official statistics, including services for training and capacity building. The platform would be gradually built according to the road map described in the background document.

5. Conclusions

Official statistics has come a long way in a short period in relation to the use of Big Data for official statistics. When the GWG was established in 2014 Big Data was sometimes mentioned as a threat to official statistics. Something that might make official statistics obsolete.

This is no longer the case.

Big Data sources has in a sense been ‘normalized’. It is no longer seen as a threat but as an opportunity. Many NSO’s have experiences in the use of Big Data for official statistics. This development will continue.

There are many challenges in the use of Big Data for official statistics – access to data being one of them.

With the Global Platform and the development of the platform the GWG will continue to strives to make partnership and capacity building easier for official statistics worldwide.

[i] Questions and comments should be directed to Niels Ploug at npl@dst.dk