Global Working Group on Big Data for Official Statistics
Task Team on Cross-cutting issues, Classifications, Frameworks and Taxonomy (TTCC)

Scoping Paper for Maintenance and Update of the Repository of Big Data projects

Table of Contents:

Action:

Make an inventory of existing and ongoing projects using Big Data for official statistics.

Responsible persons:

- Nancy Snyder (UNSD) and Kenneth Iversen (UNSD) for compiling list of Big Data projects.
- Clarence Lio (UNSD) for designing website.
- Cross-cutting issues Task Team members (Albrecht Wirthmann, Paolo Righi, Peter Struijs, Bogomil Kovachev) for discussing objective; intended audience; coverage and structure of inventory; confidentiality issues; maintenance and timeline.

Objective and intended audience:

The objective of this activity is to update the previous work of the UNSD/UNECE to maintain the existing repository on projects related to Big Data and official statistics and to include information related to quality and methodology.

The inventory will allow users to interactively search the attributes of the projects, including methodologies used in the project, quality frameworks employed, means of access to the Big Data source used, statistical domain for which Big Data was used, etc.

The TTCC should determine who the intended audience is for the inventory. The audience will in part determine the structure of the inventory and how to manage confidentiality issues. Below are the possible different audiences and the related issues:

1. <u>The Global Working Group on Big Data (GWG)</u> - If the intended audience is solely the Global Working Group on Big Data, the inventory can be maintained in one of a variety of interactive, searchable formats (perhaps Excel), either on-line on a password-protected website or on Trello. Confidentiality issues (see section VI) would be minimal, as the GWG is already a rather closed group and the members already have access to many of the projects, via the 2015 survey results or otherwise.

2. <u>National statistical offices (NSOs)</u> – If NSOs are the intended audience, the inventory should likely be maintained in a password-protected website, once the projects have been approved for dissemination by the initiating agencies to assure confidentiality concerns have been met (see section VI).

3. <u>General public</u> – If the intended audience is the general public, the number of projects with detailed information will be limited to those which are not confidential. The other projects can be partially shown or displayed in an aggregate form, based on the requests of the institutions conducting the projects.

Coverage and definitions:

The inventory will cover all completed, pilot and on-going projects using Big Data for official statistics, including those that currently exist in the UNSD/UNECE inventory, the projects submitted in the 2015 Global Survey on Big Data, and research and submissions conducted by other task team members that identify other projects using Big Data for official statistics. The inventory will be merged with a similar inventory being built by the Task Team on SDGs including projects using Big Data to monitor SDGs.

The inventory will contain the following attributes (*please note: the level of detail on these attributes may be subject to confidentiality):

1. **<u>Big Data source</u>** –
   Once all of the existing projects have been compiled, the types of Big Data sources could be analysed to determine if there are five or six main types of Big Data most commonly used in the projects. If the inventory is intended to be placed on-line and open to the public, it would be more user-friendly to have five to six main Big Data sources rather than exhaustive list. However, if the inventory is to be more of a detailed repository maintained for purposes of the GWG, more data source types can be maintained.

   Moreover, the classification/naming system of the sources should at least generally conform to the final version of the new Classification of Big Data, Deliverable 1 of the TTCC. Until that revised classification is available, the proposed sources, which are (partially) based on the existing UNECE Classification of Big Data sources, are as follows, with the main types listed first, followed by auxiliary types if needed:

   a. **Social Media**: including, but not limited to, Facebook, Twitter, blogs and comments, personal documents, Pictures: (Instagram, Flickr, Picasa etc.), videos (Youtube, etc.), internet searches, mobile text messages, e-mail.
   b. **Mobile phone location sensors**
   c. **Satellite images**
   d. **Business Data**: including, but not limited to, commercial transactions, banking/stock records, e-commerce, credit cards
   e. **Traffic sensors/webcams**
   f. **Weather and pollution sensors**

   Auxiliary data source types:
   g. <u>Other fixed sensors</u>: including, but not limited to, home automation, scientific sensors, security/surveillance videos/images
   h. <u>Administrative Data</u>: including, but not limited to, data produced by Public Agencies, medical records,

- **i.** Other mobile sensors (tracking): including, but not limited to, cars
- **j.** Data from computer systems: logs and web logs

2. **Project title or identifier**: A brief title that characterizes the objective of the project. For example, "Feasibility study on web scraping for labour market indicators".

3. **Name, country, and/or type of agency or institution that initiated the project**

4. **Area of official statistics**: identification of official statistical domain(s) with which the project is related.

5. **Applicable to SDGs monitoring**: Yes/No. If yes, the applicable SDG goal number, target and/or indicator will be included.

6. **Phase of project:**
   1) Exploration
   2) Scientific / research
   3) Pilot intended to go to production
   4) In use for production of official statistics

7. **Indication if a quality framework was applied**: Yes/No

The project titles can be hyperlinked to more detailed information page about the project, if available, including these possible topics:

- more detailed description of the objective of the project;
- how the Big Data source was accessed;
- outline of the methodological steps/procedures employed;
- brief description of the quality frameworks used;
- links to the documentation on the project from the original source.

Structure and format:

The inventory should enable user-friendly filtering and sorting by any attribute. For example, a user should be able to look at projects using a specific data source, or look at projects related to a specific statistical domain. The default setting of the inventory would be to group the projects according to Big Data source.

Confidentiailty:

For the previous Big Data survey condcuted in 2013 the respondents identified if the project(s) cited could be: 1) made public, 2) shared on a password-protected site; or 3) shared in an aggregated form. Survey respondents were *not* asked this question on the 2015 survey. Therefore, UNSD will contact each agency/institution to explain the purpose of the inventory and to request written permission to publish the aforementioned attributes of the project, and at which level of detail the projects can be published.

Updating and Maintenance:

Once permission has been received from the agencies/institutions to publish the attributes of their Big Data projects, UNSD will request a contact name and e-mail address and will explain that an automatic e-mail will be periodically sent to the contact person requesting any updates or changes to the informaiton in the inventory. A proposed time-frame for this e-mail reminder would be bi-annually, at least for the first two years. Subsequently, USND can evaluate the web traffic of the inventory web page to determine the maintenance cycle thereafter.

Sources:

This inventory will combine a number of different inventories, in particular:

- World Bank list
- UNECE/UNSD Big Data project lists
- UN Division for Sustainable Development list
- Results of 2015 Global Survey on Big Data
- Input from other Task Teams

Proposed Timeline (subject to revision):

- Mid-October 2015: finalize a Scope Paper for the inventory
- Mid-October 2015: discuss issues regarding coverage, structure, confidentiality and maintenance at the Big Data conference in Abu Dhabi
- End-October 2015: propose a preliminary structure and format for the inventory
- End-October 2015: begin contacting institutions to request written permission to publish attributes of their Big Data projects on-line
- Beginning-November 2015: finish compiling existing project lists
- End-November 2015: Finalize decisions regarding coverage, structure, confidentiality and maintenance
- Mid-December 2015: complete the first draft version of the inventory based on available project lists
- January 2016 onwards: populate on-line inventory with projects for which we have received permission to publish
- January 2016 onwards: perform regular updating and maintenance