Update and Maintenance of the Big Data Project Repository

Deliverable 1, Task Team on Cross-cutting issues, Classifications, Frameworks and Taxonomy

Ronald Jansen

United Nations Statistics Division

BigData@un.org

@UNBigData

#UNBigData2015

Big Data Project Repository

- ✓ The repository will include all on-going, pilot, and completed projects using Big Data for official statistics
- ✓ Will include projects that can be used to monitor SDGs as identified by Task Team on Big Data and SDGs
- ✓ It will build upon existing project inventories:
  - ❑ **World Bank list**
  - ❑ **UNECE/UNSD Big Data project lists**
  - ❑ **UN Division for Sustainable Development list [Paris21]**
  - ❑ **Results of 2015 Global Survey on Big Data**
- ✓ To include on-going research from other Task Teams

Big Data Project Repository

Who is the intended audience?

- The UN Global Working Group on Big Data
- National Statistical System institutes
- General public?

What is the intended use?

- User friendly, general research to see what is being done – with contact information
- Specific information regarding Methodology, Quality aspects, IT tools, Partnership models, etc.

Proposed attributes to include

- ✓ Type of Big Data Source
- ✓ Project title: a general characterization (e.g., "Feasibility study on web scraping for labour market indicators")
- ✓ Name, country, and/or type of institution that initiated the project
- ✓ Area of official statistics with which the project is related
- ✓ Applicable to SDGs monitoring: if yes, the applicable SDG goal(s)and target(s)
- ✓ Phase of project:
  - ❑ **Exploration**
  - ❑ **Scientific / research**
  - ❑ **Pilot intended to go to production**
  - ❑ **In use for production of official statistics**
- ✓ Indication if quality assessments
- ✓ Projects can be hyperlinked to more detailed info pages including links to original documentation, how the Big Data source was accessed, detailed methodology, etc.

Types of Big Data Source

- ✓ Conform to the final version of the new Classification of Big Data, deliverable of the Task Force on Cross-Cutting Issues.
- ✓ Keep the repository user-friendly, possibly limit types to the most common 5-6, based on UNECE Classification, such as:

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

Structure/Format

- 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.

Maintenance

- ✓ UNSD will **need assistance from the Big Data projects custodians** to update the information
- ✓ UNSD will request a contact name and e-mail address
- ✓ Send a periodic automatic e-mail requesting any updates or changes
- ✓ Propose to send the reminder twice per year, at least for the first two years. Thereafter, USND can evaluate the web traffic of the inventory web page to determine the maintenance cycle

Confidentiality

- ✓ **UNSD will contact institutions to request permission to publish details** of Big Data projects submitted in the 2015 Big Data survey
- ✓ Ask for similar level of confidentiality that was asked on previous 2013 Big Data survey, on whether the project can be:
  - **made public**
  - **shared on a password-protected site**
  - **shared in an aggregated form**