Governance and Management of the GWG Big Data Inventory

August 16, 2016

The GWG Big Data Inventory is a catalog of Big Data projects that are relevant for official statistics, SDG indicators and other statistics needed for decision-making on public policies, as well as for management and monitoring of public sector programs/projects. This inventory is a joint product of the World Bank and the United Nations Statistics Division (UNSD) put together on behalf of the UN Global Working Group (GWG) on Big Data for Official Statistics. The tasks related to the content of the inventory are led by the World Bank and UNSD, and the technical side is serviced by the UNSD technical team.

1. Initial version of the Inventory

The initial version of the inventory (available online at: http://unstats.un.org/bigdata/inventory/) includes project information collected by the World Bank and the UNSD through various means (surveys, direct contacts and other). The database includes 83 fields, as shown in Annex 1.

For any new addition of a Big Data project to the inventory, permission to publish needs to be obtained from the custodian of that project, before entering the project information into the existing fields of the Excel database as outlined in Annex 1 or by entering the information in the Google form. The current version of the Excel file is on the Trello Board for this community for reference.

UNSD will review the master Excel file which will then be converted to a JSON file and then sent to the UNSD web team for posting to the website.

For new projects to be included, the mandatory fields that are required are the following:

- Identification of the country, countries, or geographic region in which the project is focused
- Name and type of the institution
- Contact Information of the submitter (name and e-mail address mandatory)
- Title of the project
- Brief description of the project
- Types of data source(s) used in the project (e.g., mobile phone, satellite imagery, social media, etc.)
- The objective(s) of the project (e.g., exploration, scientific/research, pilot intended to go to production to replace existing data, etc.)
- The area(s) of official statistics that are relevant to the project

2. Adding new entries to the Inventory

Any organization or individual will be able to propose new entries to the Inventory by filling out the Google form linked to the Inventory website (see Annex 2 and available online here:

https://docs.google.com/forms/d/1Yw6wPLVjpRPeH_ribZD5hEj7vD21llwv9g53kdE_YQI/edit?ts=577fea72&pli=1). As indicated in the Google form by asterisks, the following fields are mandatory for any new projects to be added:

- Consent to publish the project on the Inventory website
- Identification of the country, countries, or geographic region in which the project is focused
- Name and type of the institution
- Contact Information of the submitter (name and e-mail address mandatory)
- Title of the project
- Brief description of the project
- Types of data source(s) used in the project (e.g., mobile phone, satellite imagery, social media, etc.)
- The objective(s) of the project (e.g., exploration, scientific/research, pilot intended to go to production to replace existing data, etc.)
- The area(s) of official statistics that are relevant to the project

The World Bank and UNSD focal points will review new proposed entries once a month and approve those that fit the criteria. The IT teams of the World Bank and UNSD will work on automating the process of uploading approved projects submitted on the Google Form into the format needed for the online database (i.e., transfer to the Excel file and conversion into a JSON file).

3. New text to webpages and minor editing of the Inventory

The World Bank and UNSD focal points will be responsible for suggesting minor changes to the content of the Inventory and for minor styling/formatting changes. UNSD will inform the UNSD web team, who will implement these changes as soon as possible.

4. More substantial changes to the Inventory

The World Bank and UNSD focal points can make suggestions on more substantial changes to the website on Trello, where all GWG members and the UNSD web team can agree/finalize suggested changes, comment on their feasibility, and provide feedback on a timeframe for how long such changes can be implemented. Such substantial changes would include the addition or editing of the basic architecture of the Inventory or its features or including a new search filter on the existing database structure.

5. Official Launch of the Inventory

The "unofficial" version of the Inventory has been launched, with 192 initial projects, available at: http://unstats.un.org/bigdata/inventory/

The aim is to launch the "official" Inventory (linked directly from the UN GWG Big Data homepage, at: http://unstats.un.org/unsd/bigdata/inventory/ before the 3rd International Conference in Dublin, on August 30-September 1, 2016.

Annex 1
Data fields in Inventory Database

Field Name in database	Field Description
ID	ID # for each project
country/Area	Country, countries or areas involved in the project
iso	ISO code (if only 1 country)
IDCountry/Regional	Flag to indicate whether the geographic area is a country or a region
incomeLevel	Income level of country (or countries, if applicable)
region	Region
organization	Organization
typeOfInstitution	Type of institution
division	Division of organization submitting the project
contactName	Contact name
contactEmail	Contact email
projectTitle	Title of the project (edited if necessary)
projectDescription	Brief description of the project (edited if necessary)
projectTitle2	Secondary title of the project (unedited version)
projectDescription2	Brief description of the project (unedited version)
bigDataSource__001	Big Data Source used (1st)
bigDataSource__002	Big Data Source used (2nd)
projectObjective__001	Project objective (1st)
projectObjective__002	Project objective (2nd)
projectObjective__003	Project objective (3rd)
projectObjective__004	Project objective (4th)
statisticsArea__001	Area of official statistics relevant to project (1)
statisticsArea__002	Area of official statistics relevant to project (2)
statisticsArea__003	Area of official statistics relevant to project (3)
statisticsArea__004	Area of official statistics relevant to project (4)
statisticsArea__005	Area of official statistics relevant to project (5)
dataProviders__001	Data Provider (1)
dataProviders__002	Data Provider (2)
otherPartners__001	Other Partner (1)
otherPartners__002	Other Partner (2)
partnershipComments	Comments on partners
intermediary	Do you use an intermediary to obtain the data?
intermediaryComments	Comment on use of an intermediary
dataAccessRights	Do you have access to the data only for this project and its purpose?
dataAccessComments	Comments on data access
coveragePeriod	Time period covered by the data
dataCoverage	Coverage and frequency of the data

frequencyComments	Comments on frequency and coverage of the data
coverageGeoPop	Coverage (geographic coverage and coverage of the population)
coverageGeoComments	Comments on geographic and population coverage
costImplication	Cost implication of the data source
costComments	Comments on cost implication
validationWithTrainingData	Are you using “training” or “ground truth” data to validate the results obtained?
validationComments	Comments on “training” or “ground truth” data validation
qualityFramework	Have any existing quality frameworks been applied to this data source
qualityFrameworkComments	Comments on quality frameworks
dataQualityConcerns	Do you have concerns about the quality of the data you have obtained?
dataQualityConcernsComments	Comments on data quality concerns
qualityAspectsEvaluated__001	Which quality aspects were evaluated (1)
qualityAspectsEvaluated__002	Which quality aspects were evaluated (2)
qualityAspectsEvaluated__003	Which quality aspects were evaluated (3)
qualityAspectsEvaluated__004	Which quality aspects were evaluated (4)
qualityAspectsEvaluated__005	Which quality aspects were evaluated (5)
qualityAspectsEvaluated__006	Which quality aspects were evaluated (6)
qualityAspectsEvaluated__007	Which quality aspects were evaluated (7)
qualityAssessmentComments	Comments on quality assessments
developedNewMethods	Have you developed new estimation methods or a methodological framework specifically related to the data source(s) used in this project?
estimationMethodologicalFrameworkComments	Comments on estimation methods or methodological frameworks
methodsUsed__001	What methods were used (1)
methodsUsed__002	What methods were used (2)
methodsUsed__003	What methods were used (3)
methodsUsed__004	What methods were used (4)
methodsComments	Comments on methods used
technologies__001	What technologies and tools were used or are being used (1)
technologies__002	What technologies and tools were used or are being used (2)
technologies__003	What technologies and tools were used or are being used (3)
technologies__004	What technologies and tools were used or are being used (4)
technologies__005	What technologies and tools were used or are being used (5)
technologiesComments	Comments on technologies used
TimeFrameToProduceIndicator	Time frame to produce the indicator
projectOutcomes	Outcomes or intended outcomes
projectPublications	Did you publish any research articles arising from your project and (separately) publications that have been very influential in your project design.
projectURL	Website of the project
publicationsComments	Comments on publications
documentation	Other documentation

sdgRelevance	Flag for applicability to SDGs (Yes/No)
sdgGoal__001	SDG goal applicable (1)
sdgGoal__002	SDG goal applicable (2)
sdgGoal__003	SDG goal applicable (3)
sdgGoal__004	SDG goal applicable (4)
sdgGoal__005	SDG goal applicable (5)
sdgGoal__006	SDG goal applicable (6)
sdgComments	SDG targets applicable

Annex 2
Google Form

https://docs.google.com/forms/d/1Yw6wPLVjpRPeH_ribZD5hEj7vD21llwv9g53kdE_YQI/edit?ts=577fea72&pli=1

APPLICATION FORM TO SUBMIT A PROJECT TO THE UN GWG BIG DATA PROJECT INVENTORY

If you are working on a project that you would like to be considered for inclusion in this Inventory, even if the project is in an initial phase, please fill out this application form. Please note that the project should either use Big Data sources and/or utilize Big Data techniques, and ideally have some relevance or implications for official statistics, SDG indicators or other statistics needed for decision-making on public policies. The Global Working Group will review submissions and include those projects that meet these criteria, or possibly contact you for further information. Please note that the information submitted below, once approved, will be made public on the GWG Big Data Project Inventory website.

*1. Do you consent to the information provided being made public on the UN GWG Big Data Project Inventory website?
Yes
No

1.1 If you have comments for the 1st question, please enter them here:

*2. Please enter the country, countries, or geographic region in which your project is focused:

*3. Name of your institution/organization.

*4. Choose the type of your institution:
Mark only one:
National statistical office
Other national governmental office
International organization
Regional organization
Academic institution
Research institute
Private sector
Other

4.1 If you selected "other", please specify:

5. Name of your division within institution:

*6. Contact name: *

*7. Contact email address: *

*8. Title of your Big Data project *

*9. Please briefly describe your Big Data project: *

*10. Please choose the type of data source(s) used in your project: *
Check all that apply.
Mobile phone data
Satellite imagery or aerial imagery data
Social media data

Credit card data
Smart meter electricity data
Road sensor data
Public transport usage data
Ships identification data
Health records
Web scraping data
Scanner data
Other

10.1 If you selected "other", please specify:

*11. Please choose the objective(s) of your project:
Check all that apply.
Exploration
Scientific / research
Pilot intended to go to production to improve timeliness
Pilot intended to go to production to supplement existing data
Pilot intended to go to production to replace existing data
For the production of statistics
Other

12. Please choose the area(s) of official statistics that are relevant to your project: *
Check all that apply.
Agricultural statistics
Business statistics
Crime statistics
Demographic and social statistics
Economic and financial statistics
Energy statistics
Environmental statistics
Geospatial statistics
Governance statistics
Information society / ICT statistics
Labour statistics
Mobility statistics
Price statistics
Tourism statistics
Transportation statistics
Vital and civil registration statistics
Other

12.1 If you selected "other", please specify:

13. Please choose the provider(s) of Big Data for your project:
Check all that apply.
Mobile phone operator
Satellite or aerial imagery provider
Social media provider
Intermediary Big Data provider
Cloud server provider

14. What other types of partner(s) are involved in your project:
Check all that apply.
Technology
Partner

Research or Academic institute
Government institute
International organization
Other.

14.1 Comments about access or partnerships:

15. Do you use an intermediary (company or research institute) to obtain and prepare the data for your project?
Mark only one.
Yes
No

15.1 Please explain:

16. Do you have access to the data only for this project and its purpose?
Mark only one.
Only for this project
Broader access rights

16.1 Please explain, if you selected "broader access rights":

17. Time period covered by the data:

18. Coverage and frequency of the data
Mark only one.
Only a portion of all data
All available data
Other

18.1 Please explain, if you selected "other":

19. Coverage (geographic and population) of the data:
Mark only one.
Part of country / high % of market
Whole country / high % of market

19.1 Please explain as needed:

20. Cost implication:
Mark only one.
Free
Commercial

20.1 Please explain:

21. Are you using “training” or “ground truth” data to validate the results obtained?
Mark only one.
Yes
No

21.1 Please explain the validation process:

22. Have any existing quality frameworks been applied to this source, in terms of:
Check all that apply.
Quality of processing/throughput

Quality of output statistics
Quality of source/input
Quality of processing/throughput

22.1 Please explain if evaluation is both qualitative and quantitative:

23. Do you have concerns about the quality of the data you have obtained?
Mark only one.
Yes
No

23.1 Please explain what measures you took or might take to address them:

* 24. In the overall evaluation of your Big Data project do you evaluate the following quality aspects? *
Check all that apply.
Institutional/Business
Environment
Privacy and Security
Completeness, Usability, Time Factors
Accuracy, including selectivity
Coherence, including linkability to other sources
Validity
Accessibility, Relevance
Other
None

24.1 If you selected "other", please specify:

25. Have you developed new estimation methods or a methodological framework specifically related to the data source(s) used in this project?
Mark only one.
Yes
No

25.1 Please explain:

26. In this project, what methods were used or are being used during the Big Data processing life cycle?
Check all that apply.
Machine learning (Random forest, etc.)
Supervised learning
Bayesian techniques
Neural networks
Decision Trees
Data visualization methods
Traditional statistical methods
Other methods

26.1 If you selected "other", please specify:

27. In this project, what technologies and tools were used or are being used during the Big Data processing life cycle?
Check all that apply.
Hadoop Clusters
RHadoop
Spark
SAS Visual Analytics

NoSQL database
Column store database
Relational database
Spreadsheet
GIS
Cloud services
Data mining tools
Data visualization tools
Other

27.1 If you selected "other", please specify:

28. Time frame to produce the indicator:

29. Please comment on the outcomes or intended outcomes of this Big Data project:

30. Did you publish any research articles arising from your project and (separately) publications that have been very influential in your project design?
Mark only one.
Yes
No

30.1 If yes, please provide link(s) to publication(s) or the project website, if applicable, below.
We encourage you to send us documentation prepared by your office on the topic of Big Data.
Please send it to bigdata@un.org.

30.2 Please explain if necessary:

31. Is this project relevant for compiling and/or supporting the measurement of SDG indicators?
Mark only one.
Yes
No

32. For which Sustainable Development Goals (SDGs) does this project have implications?
Check all that apply.
1 No Poverty
2 Zero Hunger
3 Good Health & WellBeing
4 Quality Education
5 Gender Equality
6 Clean Water & Sanitation
7 Affordable & Clean Energy
8 Decent Work & Economic Growth
9 Industry, Innovation & Infrastructure
10 Reduced Inequalities
11 Sustainable Cities & Communities
12 Responsible Consumption & Production
13 Climate Action
13 – Life below Water
15 Life on Land
16 Peace, Justice & Strong Institutions
17 Partnerships for the Goals

33. For which SDG indicator(s) (please specify by number) does this project have implications? Or provide any other comments relevant to SDGs below.