

# **Introduction to Models and Standards for Statistical Modernisation**

## **Session 3 : Big Data and other alternative data**

**Building national capacities on  
International Standards for Official Statistics  
Chisinau 19-21 September, 2022**

# CONTENTS



1. What is Big Data?
2. Types of Big Data
3. Why use Big Data
4. Challenges of using Big Data
5. Implications for an NSO
6. Other alternative data sources
7. Examples

# 1. WHAT IS BIG DATA?

*Big data are data sources that can be – generally – described as: “high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making.”*

Gartner

- The term Big Data refers in general to data generated by business transactions, social media, phone logs, communication devices, web scraping, sensors etc.
- Big data is often largely unstructured, meaning that it has no pre-defined data model and/or does not fit well into conventional relational databases
- Can involve terabytes, petabytes and even exabytes of data created and collected over time.

## 2. TYPES OF BIG DATA?

1. **Transactional data** is generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts.
2. **Sensors/meters and activity records from electronic devices:** the quality of this kind of source depends mostly on the sensor's capacity to take accurate measurements in the way it is expected. Machine data is defined as information generated by industrial equipment, sensors installed in machinery, and even web logs that track user behaviour.
3. **Social interactions:** this covers data produced by human interactions through a network. The most common is the data produced in social networks.
4. **Citizen-generated data (CGD)** is data produced by non-state actors under the active consent and participation of citizens to primarily monitor, demand or drive change on issues that affect them directly.

UNECE classification of Big Data sources:

<https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data>

### 3. WHY USE BIG DATA?

- Big Data has attracted much from NSOs regarding the potential to complement traditional statistics.
- This is particularly important in the context of the need to measure and monitor progress towards SDGs.
- Big Data has the potential to supplement, complement or partially replace, existing statistical sources.
- It could also improve estimates or generate completely new statistical information in a given statistical domain or across domains.

[“What does Big Data mean for official statistics?”](#)

## 4. CHALLENGES USING BIG DATA

- Big Data is widely used in the commercial sector for business analytics
- There is less evidence of its use thus far in the world of official statistics....
- The technology needed to process these huge data sets is available and maturing, the biggest obstacle for an NSO is often to gain access to the data.
- Private sector may take advantage of the Big data era and produce more and more statistics that beat official statistics on timeliness and relevance



## 4. CHALLENGES USING BIG DATA



- Private companies know the value of data and don't want to give it away...
- Legal constraints concerning privacy and confidentiality of client information;
- The costs of setting up the necessary infrastructure and training staff for a non-core business-related activity.

## 4. CHALLENGES IN USING BIG DATA

**Data privacy** - SOs need to ensure that data sources and indicators used are obtained without violating privacy or confidentiality regimes.

**Costs** - NSOs must also invest in security layers and adapt traditional information technology techniques such as cryptography, anonymisation, and user access control to Big Data characteristics.

### **Data quality:**

- Big Data is often largely unstructured.
- Diverse structures coming from diverse sources in various formats such as logs, call-centres, web scrapes and social media.
- Data formats will differ, and matching them can be problematic.

### **Unreliable data:**

- not always created using rigorous validation methods which can adversely affect quality.
- Can be inaccurate, contain wrong information, contain duplications and other contradictions.





## 4. CHALLENGES USING BIG DATA

### **Methods:**

- using Big Data requires new methods and techniques.

### **Data impermanence:**

- an NSO cannot guarantee that a data source will be reliable as it has no control over or relationship with the data owner as with traditional data sources.
- Formats can change at any time without warning that can render data capture and subsequent processes that have been put in place by the NSO unworkable.
- Data sources can even disappear completely if the business rules generating the data are changed.

### **Data gaps:**

- SDGs have a fundamental commitment to leave no-one behind.
- Vulnerable populations may not be covered by Big Data if using sources such as mobile phones are not available to the poorest and most marginalised groups of society.



## 5. IMPLICATIONS FOR AN NSO



- Collaboration of NSOs with private data source owners.
- Statisticians are needed with a different mind-set and new skills.
- NSOs need to develop the necessary internal analytical capability through specialised training.
- NSO as brokers of Big Data?
- NSO to provide “Quality Stamp” ?


## 6. EXAMPLES

A screenshot of a web browser displaying the "Big Data Inventory" page on statswiki.unece.org. The browser's address bar shows the URL "https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Invent...". The page has a green header with a "Spaces" dropdown menu. Below the header, there is a section titled "Big Data Inventory" with a colorful elephant icon. Underneath, a "PAGE TREE" section lists various statistical projects and studies from different countries and organizations, including Australia, Poland, ESCAP, Eurostat, Italy, Mexico, China, Romania, and Canada.

← → ↻ 🏠 🔒 https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Invent...

📧 Mail 📄 Online Card Service... 🏠 NatWest 🌐 BNP 📄 Espace client Gener... 🔗 LinkedIn 🗨️

📄 Spaces ▾

 Big Data Inventory

PAGE TREE

- Australia (ABS) - Social Linked (semantic) Data Processing for Various Statistical Use
- Central Statistical Office of Poland - Estimating demand on labour market by analy:
- ESCAP - Developing a Curriculum and Training Modules on Using Big Data for Offic
- Eurostat - Feasibility study on the use of mobile positioning data for tourism statist
- Eurostat - Multi-purpose consumer price statistics, sub-project Scanner Data/Web :
- Italy (Istat) - Internet as a Data Source for ICT Usage by Enterprises and Public Insti
- Italy (Istat) - Persons and Places: Mobility Estimates based on Mobile Phone Data
- Italy (Istat) - Specific purpose geographic basins and population statistics using mc
- Italy (Istat) - Use of scanner data for consumer price index
- Mexico (INEGI) - Tweet Analysis
- National Bureau of Statistics of China - Big Data Enterprise Statistical Indicator Ten-
- National Bureau of Statistics of China - Online Price Changes of Means of Productic
- Romania National Statistical Institute (INS) - Using scanner data
- Statistics Canada - Non-Residential Buildings Inventory: Feasibility Study

<https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Inventory+Home>

# 6. EXAMPLES



<https://thedataeconomylab.com/explorer/big-data-project-inventory/>



Our Work **Stewardship Explorer** Tools and Guides Videos Resources

## Big Data Project Inventory

PUBLIC CIVIC

### DESCRIPTION

The GWG Big Data Inventory is a catalog of Big Data projects that are relevant for official statistics, SDG indicators and other statistics. 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.



### MODEL TYPE

Open Data

### INNOVATED BY

Non-Governmental

### PLACE OF INNOVATION

# 6. EXAMPLES



Home > Search results for: 'big data'

## Filter your results

Big Data Source 


- Satellite imagery or aerial imagery data (8)
- Mobile phone data (17)
- Web scraping data (24)
- Smart meter electricity data (5)
- Scanner data (21)
- Other (26)
- Credit card data (3)
- Road sensor data (5)
- Ships identification data (2)

## Search results for: 'big data'

Items 1-10 of 209

### Related search terms

- [Data sources](#)
- [Big data UN SDGs](#)
- [Big Data for Financial Inclusion](#)
- [Big Data UN GWG](#)
- [big data project](#)

Name Z-A 

Feedback 

<https://marketplace.officialstatistics.org/catalogsearch/result/?q=big+data>

## 7. OTHER ALTERNATIVE DATA SOURCES



*“Synthetic data generated from computer simulations or algorithms provides an inexpensive alternative to real-world data that’s increasingly used to create accurate AI models.”*

# HLG-MOS Synthetic Data Test-Drive

LIBRARY OF RESOURCES COMPILED FROM  
HLG-MOS CHALLENGE 2022

The test-drive contains an archive of collective expert knowledge gathered from the participant submissions for the High Level Group on the Modernization of Official Statistics (HLG-MOS) Synthetic Data Challenge, held in January 2022. This platform aims to offer some insights from the members of National Statistical Organizations (NSOs) and the synthetic data community at large, regarding the subject matter expertise in synthetic data generation, and their perspective towards utility and privacy of synthesized data. The High-Level Working Group for the Modernisation of Statistics ([HLG-MOS](#)) is a function of the United Nations Economic Commission for Europe ([UNECE](#)). The HLG-MOS Challenge used NIST's [SDNIST: Synthetic Data Benchmarking Library](#) as an evaluation tool. These results are shared with permission under NIST agreement DTA-22-011. These data are for informational



## 8. RESOURCES



- UN Big Data inventory:  
<https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Inventory+Home>
- GWG Big Data project inventory:  
<https://thedataeconomylab.com/explorer/big-data-project-inventory/>
- Big data in official statistics, CBS (2020); <https://www.cbs.nl/en-gb/background/2020/04/big-data-in-official-statistics>
- Recommendations for access to data from private organizations for official statistics, Global Working Group on Big Data for Official Statistics (2016);  
[https://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%202%20\(i\)%20a%20-%20Recommendations%20for%20access%20to%20data%20from%20private%20organizations%20for%20official%20statistics%20Draft%2014%20July%202016.pdf](https://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%202%20(i)%20a%20-%20Recommendations%20for%20access%20to%20data%20from%20private%20organizations%20for%20official%20statistics%20Draft%2014%20July%202016.pdf)
- UN Global Working Group on Big Data.  
<https://unstats.un.org/bigdata/index.cshtml>
- Synthetic Data for Official Statistics: A Starter Guide  
[https://github.com/usnistgov/HLG-MOS\\_Synthetic\\_Data\\_Test\\_Drive/raw/nist-pages/guide/Synthetic%20Data%20for%20NSOs%20A%20starter%20guide.pdf](https://github.com/usnistgov/HLG-MOS_Synthetic_Data_Test_Drive/raw/nist-pages/guide/Synthetic%20Data%20for%20NSOs%20A%20starter%20guide.pdf)



Thank you for your attention....

Q&A