Overview

- Big data – definition, data sources
- Statistical Data Infrastructure
- Quality framework for Big Data
Big data
Definition – data sources -

Big Data

Wikipedia:
The term has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

(https://en.wikipedia.org/wiki/Big_data)
Big Data

Wikipedia:
"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.

(https://en.wikipedia.org/wiki/Big_data)

Big Data

Wikipedia:
Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

(https://en.wikipedia.org/wiki/Big_data)
Big Data

Wikipedia:

Source: Medium (https://medium.freecodecamp.org/a-brief-history-of-serverless-or-how-i-learned-to-stop-worrying-and-start-loving-the-cloud-7b2f630331ed)


United Nations Statistics Division
What are common sources of Big Data?

- Automatically generated data in electronic format, such as mobile phone data, social media data, electronic commercial transactions, sensor networks, smart meters, GPS tracking device, or satellite images
- High frequency, and/or fine granularity, and/or wide coverage
As a "special case" of human mobility, tourism is a human activity that leaves multiple traces, as a digital footprint or captured by sensors.

[Image showing various activities related to tourism]

**Taxonomy of big data sources (Eurostat 2017)**

- **Communication systems**: Mobile network operator data, Smart mobile devices data, Social media posts
- **World Wide Web**: Web activity, Dynamic websites, Static websites
- **Business process generated data**: Flight booking systems, Hotel data, Financial transactions
- **Sensors**: Traffic loops, Smart energy meters, Vessel radio identification
- **Crowd sourcing**: Volunteered geographic information (OpenStreetMap), Wikipedia contents, Picture collections

---

United Nations Statistics Division
Why are Big Data important?

- **Big Data can keep official statistics relevant** – private sector moves fast
- **Big Data are part of modernization of statistical systems** – new production processes and partnerships
- **Big Data can help core national statistics** – for integrated economic, social and environmental policies
- **Big Data can help meeting the data demand of the 2030 agenda** – monitoring policies – “leave no one behind”
- **Big Data are needed for agile statistics** – for emergency issues

---

**Statistical Data Infrastructure**
1. Statistical data

2. Register Data
3. Other Administrative Data

United Nations Statistics Division

- Tax (VAT) data
- Social security data
- Water consumption data
- Trade associations
- Social benefits data
- Credit card or scanner data
- Smart meter or sensor data
- Satellite imagery and aerial data
- Mobile phone data
- Social media or web-scraping data
- Satellite imagery and aerial data

4. (Privately held) Big Data

United Nations Statistics Division
United Nations Statistics Division

Statistical Data Infrastructure

1. Statistical data
2. Register Data
3. Other Administrative Data
4. Big Data

Geo-Spatial integration

Statistical integration: SNA and SEEA

United Nations Statistics Division
Integrated statistics approach

Statistical operations

Outputs / Dissemination

Macroeconomic accounts

Household and demographic statistics

Economic & environmental statistics

Data integration

Data processing

Data collection

Registers and frames

Surveys

Statistical infrastructure

Inputs

Standards and methods

Information, Communication Technology (ICT)

Management and internal policy

Institutional arrangements

Institutional setting

Quality framework for big data
Framework for NSO to assess quality of big data

General approach

**Quality**: to be evaluated in the of intended use (‘fitness for use’)

**Generic statistical business process model:**

```
<table>
<thead>
<tr>
<th>Input</th>
<th>Throughput</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>acquisition</td>
<td>transformation</td>
<td>reporting</td>
</tr>
</tbody>
</table>
```

**Framework**: For each phase define appropriate quality dimensions and quality indicators

Hyper-dimensions

The concept of hyper-dimension was taken from the Netherland administrative data quality framework.

- **Source**: Related to the type of data, the entity from which the data is obtained, and how it is administered and regulated.
- **Metadata**: Description of concepts, file contents, and processes.
- **Data**: Related to quality of the data itself.
Framework for NSO to assess quality of big data

Quality dimensions

- Institutional/business environment
- Privacy and security, complexity
- Completeness, usability, time factor
- Accuracy
  - selectivity
- Coherence
  - linkability
- Validity
- Accessibility, clarity, relevance

<table>
<thead>
<tr>
<th>Hyperdimension</th>
<th>Quality Dimension</th>
<th>Factors to consider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Institutional Environment</td>
<td>Sustainability of the entity-data provider</td>
</tr>
<tr>
<td></td>
<td>Privacy and Security</td>
<td>Reliability status, transparency, interpretability</td>
</tr>
<tr>
<td>Metadata</td>
<td>Complexity</td>
<td>Technical constraints, Structured or Unstructured</td>
</tr>
<tr>
<td></td>
<td>Readability</td>
<td>Presence of hierarchies and nesting</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>Metadata is available, interpretable and complete</td>
</tr>
<tr>
<td></td>
<td>Usability</td>
<td>Resources required to import and analyse</td>
</tr>
<tr>
<td></td>
<td>Time-related</td>
<td>Timeliness, Periodicity, Changes through time</td>
</tr>
<tr>
<td></td>
<td>Linkability</td>
<td>Presence and quality of linking variables</td>
</tr>
<tr>
<td></td>
<td>Coherence</td>
<td>Use of standards</td>
</tr>
<tr>
<td></td>
<td>Validity</td>
<td>Transparency of methods and processes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Soundness of methods and processes</td>
</tr>
</tbody>
</table>

United Nations Statistics Division
Thank you!
Terima kasih!

Karoly Kovacs
United Nations Statistics Division | Department of Economic and Social Affairs
Email: bigdata@un.org
http://unstats.un.org/unsd/bigdata

United Nations Statistics Division