What has changed in official statistics since using Big Data and data science for the last 10 years?

Ronald Jansen
Assistant Director
United Nations Statistics Division
UN Committee of Experts on Big Data and Data Science for official statistics

Mandate (Decision 45/110 – 2014)

• Provide strategic vision of a global programme on Big Data for official statistics;
• Promote practical use of sources of Big Data and find solutions for
  • Methodological issues,
  • Legal issues of access to data sources;
  • Privacy issues
  • Data security issues;
  • Cost benefit analysis
• Promote capacity building
• Foster Communication and Advocacy
• Build Public Trust
UN Committee of Experts
on Big Data and Data Science for official statistics

- **Regional Hub LAC, Brazil**
- **Regional Hub A&P Indonesia/UNESCAP**
- **Regional Hub Africa, Rwanda/UNECA**
- **Regional Hub MENA, Dubai, UAE**
- **Global Hub on ARIES for SEEA, Bilbao**
- **Global Hub on Big Data & Data Science, Hangzhou**

**UN Global Platform**
- **Scientific Committee**
  - Data Science Leaders Network

**UNGP Committee**
- **CEBD Advisory Board**
- **CEBD Bureau**
- **Task Teams**

**Access to Privately-Held Data**
- **AIS Data**
- **Scanner Data**
- **Mobile Data**

**Earth Observations**
- **Privacy Enhancing Technologies**
- **Training, Skills & Capacity Building**
- **Data Science & SDG Localization**
UN Global Platform

Purpose: Global collaboration on Data innovation

Provides:

- Access to Data
- Access to Technology services
- Access to Expertise
- Making use of Big Data possible for small offices
Data

Global data
- AIS Vessel tracking data
- Satellite imagery data

Synthetic data
- Mobile phone data
- Smart surveys

Data as a Service
- Trade data
- Shipping register data
- Port activity
- Global group registers
Technology services

**Cloud Services**
- AWS
- Google Cloud Platform
- MS Azure
- Alibaba

**Developer Services**
- Jupyter Notebook
- R, Python
- Colab

**Supports Services**
- Cloudflare
- Net App
- Google suite
Expertise

Data engineers
- Cloud experts of NSO
- IT experts of NSO
- AWS support

Data scientists
- AIS experts
- Flowminder
- Positium
- Open Mined
- University of Tokyo
- Bocconi University
- University of Oslo

Statisticians
- Statistics Canada
- Statistics Netherlands
- ONS, UK
- BPS Indonesia
- UN Statistics Division
- ISTAT
- Statistics South Africa
- Statistics Poland
- World Bank
- Statistics Denmark
- NBS China
- FCSC UAE
- NISR Rwanda
- IBGE Brazil
Projects

- Port activity
- Maritime emissions
- PET lab
- ARIES for SEEA
- Sen2Agri
- SDG finance
- Social responsibility

Technology stack

- .STAT
- Semantic web
- AWS Cloud formation
- Kubernetes

Statistical Domains

- Transport / Trade
- Prices, CPI
- Tourism
- Migration
- Population dynamics
- Information society
- Displacement
- Climate Change
- Environment
- Agriculture
- SDG indicators
- Privacy protection
Mission and Strategies

The demand for more diversified, sophisticated and rapid statistical services could be met by leveraging the emerging sources of Big Data, such as those relating to remote sensing imagery, transactional and social media data and mobile device data.

Statistical agencies around the world have a strong interest in investigating the viability of using satellite imagery data to improve official statistics on a wide range of topics spanning agriculture, the...
Earth Observations for Official Statistics

Satellite Imagery and Geospatial Data Task Team report

5th December 2017
Analysis-ready data image collections

Analysis-ready data (ARD) are images that are ready for analysis without the need for further preprocessing or transformation. They simplify and accelerate the analysis of Earth observation data by providing consistent and high-quality data that are standardized across different sensors and platforms. ARD data is typically provided as a collection of tiles, where each pixel contains a single value for each spectral band for a given date.

ARD collections are available in cloud services such as Amazon Web Service, Brazil Data Cube, Digital Earth Africa, Swiss Data Cube, and Microsoft's Planetary Computer. These collections have been processed to improve multivariate comparability. Radiance measures at the top of the atmosphere were converted to ground reflectance measures. In general, the timelines of the images of an ARD collection are different. Images still contain cloudy or missing pixels; bands for the images in the collection may have different resolutions. Figure 9 shows an example of the Landsat ARD image collection.

Figure 9: ARD image collection (Source: USGS. Reproduction based on fair use doctrine).

ARD image collections are organized in spatial partitions. Sentinel-2/2A images follow the Military Grid Reference System (MGRS) tiling system, which divides the world into 60 UTM zones of 8 degrees of longitude. Each zone has blocks of 6 degrees of latitude. Blocks are split into tiles of $110 \times 110$ km$^2$ with a 10 km
Task Team on Use of Mobile Phone Data for Official Statistics

**About Mobile Phone Data**

**Introduction**

The statistical community has the obligation of exploring the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working. Use of Big Data could also support the monitoring of the Sustainable Development Goals (SDGs) by improving timeliness, frequency, detail and relevance of indicators without compromising their...
1) Tourism statistics (lead: Indonesia)
2) Migration statistics (Lead: Georgia)
3) Dynamic population (lead: Positium)
4) Transport and commuting statistics (lead: UAE)
5) Information society indicators (lead: ITU)
6) Displacement in disaster context (lead: University of Tokyo)
Data Flow

PHASE I
- Clean & Data QA
- Raw data
- Request for corrections

PHASE II
- Process data
- Raw data
- Data Model & Results
- Improved methodology
- Methodology Report

PHASE III
- Processed Data (Data Model)
- MNO-based Aggregated Results
- Aggregated Results Expanded to General Population
- Interactive Applications, Maps Animations
- Pilot Project Final Report

Global Network of Data Officers and Statisticians, April 20, 2021
**GERMAN TOURISTS** in Khasab

Large number of **trips**, but small number of **spent nights**
This is caused by cruise ships

**Salalah**

Large number of **trips** and **spent nights**
Stable number of tourists who spend the night
Task Team on Use of Scanner Data and Webscraping for Price Statistics
Prices from digital sources

Web Scraping data for:
- Clothing stores
- General Merchandisers
- Home improvement
- Electronics and Appliances

API data for:
- Airlines
- Hotels
- Car Rentals

Scanner data for:
- Food
- Personal Care
- Household operations

In-house Internet collection of:
- Travel
- Transportation
- Communications
- Furniture
- Services
Task Team on Use of AIS Shipping data for Maritime Trade and Transport Statistics

Trade Disruptions in the Red Sea
Red Sea | 16 December, 2023 – Ongoing

Event name and type: Trade disruptions in the Red Sea (near Bab-el-Mandeb Straits) due to attacks on commercial ships

Event date: Redceded traffic since 16 December, 2023 – Ongoing

Event description: Attacks on commercial ships prompted shipping companies to re-route traffic away from the Red Sea. A systematic increase in shipping time has taken place, amounting to an increase of over 22,000 transit days annually. Further details are discussed in the IMF online update.

Main economies affected: Many economies in the Middle East, Europe, Asia, and Africa rely heavily on the Red Sea shipping lane for exports and imports. It is particularly important for oil exports from the Middle East to Europe and East-West trade.

Checkpoints in affected area: Bab-el-Mandeb Straits, Suez Canal, Cape of Good Hope

Suez Canal (located at the northern entrance to the Red Sea)
Task Team on Privacy Enhancing Technologies for Official Statistics

Privacy-Enhancing Technologies

- Introduction
- Methodologies
- Case Studies
- Standards
- Legal aspects

Introduction

The Privacy-Enhancing Technologies Task Team (PETTT) is advising the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD) on Big Data on developing the data policy framework for governance and information management of the global platform, specifically around supporting privacy-enhancing techniques.

The task team has been active since April 2018 and has released the UN Privacy Preserving Techniques Handbook. This document describes motivations for privacy-preserving approaches for the statistical analysis of sensitive data; presents examples of use cases where such methods may apply; and describes relevant technical capabilities to assure privacy preservation while still allowing analysis of sensitive data.
**CASE STUDY DESCRIPTION**

**HIGH LEVEL FUNCTIONAL PERSPECTIVE**

### Hospital

<table>
<thead>
<tr>
<th>Cryptographic ID</th>
<th>Uses eHealth Coach?</th>
<th>Wellbeing Score (PHRIS)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

What is the average education level of patients using the eHealth coach vs. patients not using the coach? (*)

### Insurer

<table>
<thead>
<tr>
<th>Cryptographic ID</th>
<th>Total Medical Costs</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100000</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>6000</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

How many insures with medical costs over 56000 are using the eHealth coach? (*)

### NSO

<table>
<thead>
<tr>
<th>Cryptographic ID</th>
<th>Education Level</th>
<th>Income</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>25000</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
<td>100000</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>L</td>
<td>43000</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

What is the average wellbeing score of patients using the eHealth coach with income above 10000 vs. below 10000? (*)

*For illustration purposes only, actual allowed queries are subject to implemented smart contract business rules*

**Figure 3.13:** An example of Private Set Intersection with Analytics (PSI-A)
Developing a Privacy Preserving Record Linkage toolkit

Overview

The Office for National Statistics (ONS) - along with other public sector institutions - rely on the ability to link datasets to produce new analysis and improve statistics for decision-making. As Sir Ian Diamond, the National Statistician, says “We find ourselves living in a society which is rich with data and the opportunities that comes with this. Yet, when disconnected, this data is limited in its usefulness. ... Being able to link data will be vital for enhancing our understanding of society, driving policy change for greater public good and minimising respondent burden.”

Data linking often needs to happen across organisational and national boundaries, which can create data privacy risks as personal information such as names, addresses, and dates of birth are often needed to do this accurately. The ONS takes very seriously its responsibility to link datasets securely, ethically and robustly, and is taking a leading role in exploring how new technology can help us achieve this.

Today, we are releasing an experimental Privacy Preserving Record Linkage toolkit, which we hope will help organisations with
This approach to overlapping security has been called the “Swiss cheese” model: imagine thin layers of Emmental, stacked so that the holes in one layer are covered by the next layer. The approach is designed to take advantage of several layers of security. It uses a combination of algorithms, encryption, and secure cloud technologies that reinforce each other and ensure that sensitive information cannot be recovered by those who should not see it.

**Toolkit**

Our design focuses on minimising the amount of infrastructure configuration for data owners, as we wanted the toolkit to be as simple as possible to use. The first element of the toolkit is a Python package that implements an experimental private data linkage algorithm. The algorithm uses trainable hash embeddings to compare and match datasets. Python users can download the package and use it to try the algorithm's performance on their data.
Training in Big Data and Data Science for official statistics

- Big Data Training Curriculum
- E-Learning Courses
- Big Data Maturity Matrix
- Training of data scientist in academic centers
- Big Data Competency Framework
- Mentorship
Data Science Leaders Network

• **Automation** of the statistical production processes (increase efficiency and improve quality)

• **Supplementary indicators** produced for emerging issues to provide additional insights

• **Changing statistical production:**
  Example – webscraping of prices from the internet combined with traditional price surveys to produce regular consumer price indices
UNCEBD partners with about 100 institutes
Survey: 10-year review on use of Big Data
Big Data and Data Science for Official Statistics

10-year review
(July-August 2023)

Survey
with 87 institutes

Interviews with
41 experts
Geographical breakdown of institutes

<table>
<thead>
<tr>
<th>Region</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>14</td>
</tr>
<tr>
<td>Asia</td>
<td>21</td>
</tr>
<tr>
<td>Europe</td>
<td>30</td>
</tr>
<tr>
<td>Latin America and the Caribbean</td>
<td>9</td>
</tr>
<tr>
<td>Middle East</td>
<td>7</td>
</tr>
<tr>
<td>North America</td>
<td>2</td>
</tr>
<tr>
<td>Oceania</td>
<td>4</td>
</tr>
</tbody>
</table>
Questions for Survey and Interview

GAMSO structure of the **Survey**
(Generic activity model of statistical organizations)

- strategic vision,
- legislation,
- institutional arrangement and partnerships,
- data sources,
- methodology and quality assurance,
- communication and stakeholders’ consultations,
- human resources,
- IT management

**Categories of the interview questions**

- UNCEBD mandate and value proposition
- Terms of References of the task teams and the regional hubs
- UN Global Platform
- Communication
- Strategic Framework
- Organizational adaptability
- Use of new data sources
10. High-level goals of multi-year plan
   Please check all those topics below which are part of your innovation strategy
38. **Infrastructure**

Which of the following options describes the change of technology infrastructure in your institute for the processing of alternative data sources, like Big Data?

You can choose more than one option.
New Data Sources

20. **Alternative data sources, including Big Data or web scraping and other privately held data**
What other data sources do you use or are you considering using in the future?
Task team participation – in which task teams has your office been actively participating

### National Statistical Offices
- Task Team on use of Satellite Data for agriculture statistics
- Task Team on use of Mobile Phone data for official statistics
- Task Team on use of Scanner data and web scraping for price statistics
- Task Team on use of AIS vessel tracking data for maritime transport and trade statistics
- Task Team on the use of Privacy Enhancing Technologies for official statistics

### International Organizations
- Task Team on Training, Competencies and Capacity Development
- Task Team on Big Data for Sustainable Development Goals
- Task Team on Rural Access to All-season roads
- Task Team on Facilitating Access to privately held data
- None of these task teams
Task teams of interest – in which task teams would your office be interested to participate

**National Statistical Offices**
- Task Team on use of Satellite Data for agriculture statistics
- Task Team on use of Mobile Phone data for official statistics
- Task Team on use of Scanner data and web scraping for price statistics
- Task Team on use of AIS vessel tracking data for maritime transport and trade statistics
- Task Team on the use of Privacy Enhancing Technologies for official statistics

**International Organizations**
- Task Team on Training, Competencies and Capacity Development
- Task Team on Big Data for Sustainable Development Goals
- Task Team on Rural Access to All-season roads
- Task Team on Facilitating Access to privately held data
- None of these task teams
Almost 4 out of 5 NSOs have explicitly incorporated references to modernization, innovation, data science, and the use of big data, into their strategic agendas.

Access to private sector data together with protection of data privacy are main priorities in the innovation strategies.

Correspondingly, more than 4 out of 5 NSOs have updated or are in the process of updating their statistical legislation to facilitate access to privately held data.

About half of NSOs and IOs are actively developing data science capabilities in their institutes.
Almost 4 out of 5 NSOs have a roadmap to develop capacity in new areas, such as data science, data engineering or similar.

Whereas most NSOs have gradually upgraded their IT infrastructure, only about half using Cloud services.

About 2 out of 3 NSOs have not yet participated in the UNCEBD task teams but all are interested to join in one of the task teams.
It is recommended that the communication of UNCEBD will be improved through

- user-friendly upgrades and consistent content updates of the UNCEBD website,
- improving the organization of the international conferences, and
- streamlining communication tools, such as newsletters, social media, and more focused content.
The mandate of UNCEBD is to provide strategic vision, direction and coordination for a global programme on the use of data science, Big Data and other alternative data sources for official statistics. Within this global program UNCEBD should:

- conduct use cases, while facilitating data access and protecting data privacy;
- develop solutions for many methodological, technical and legal challenges;
- promote capacity-building activities;
- promote partnerships with private sector and academia;
- promote the integration of statistical and geospatial information;
- develop communication strategies to maintain public trust.
What is next?

- More Data Science and AI
- More Partnerships across the Data Landscape
- Data Governance
  - Equal access to data
  - Privacy enhancing technologies & improved data sharing & risk/cost of not-sharing
  - Information integrity & data ethics
8th International Conference on Big Data and Data Science for official statistics

Informing Climate Change and Sustainable Development policies with integrated data

Venue: Euskalduna Conference Centre,
Bilbao, Spain
Dates: 10 to 14 June 2024