



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

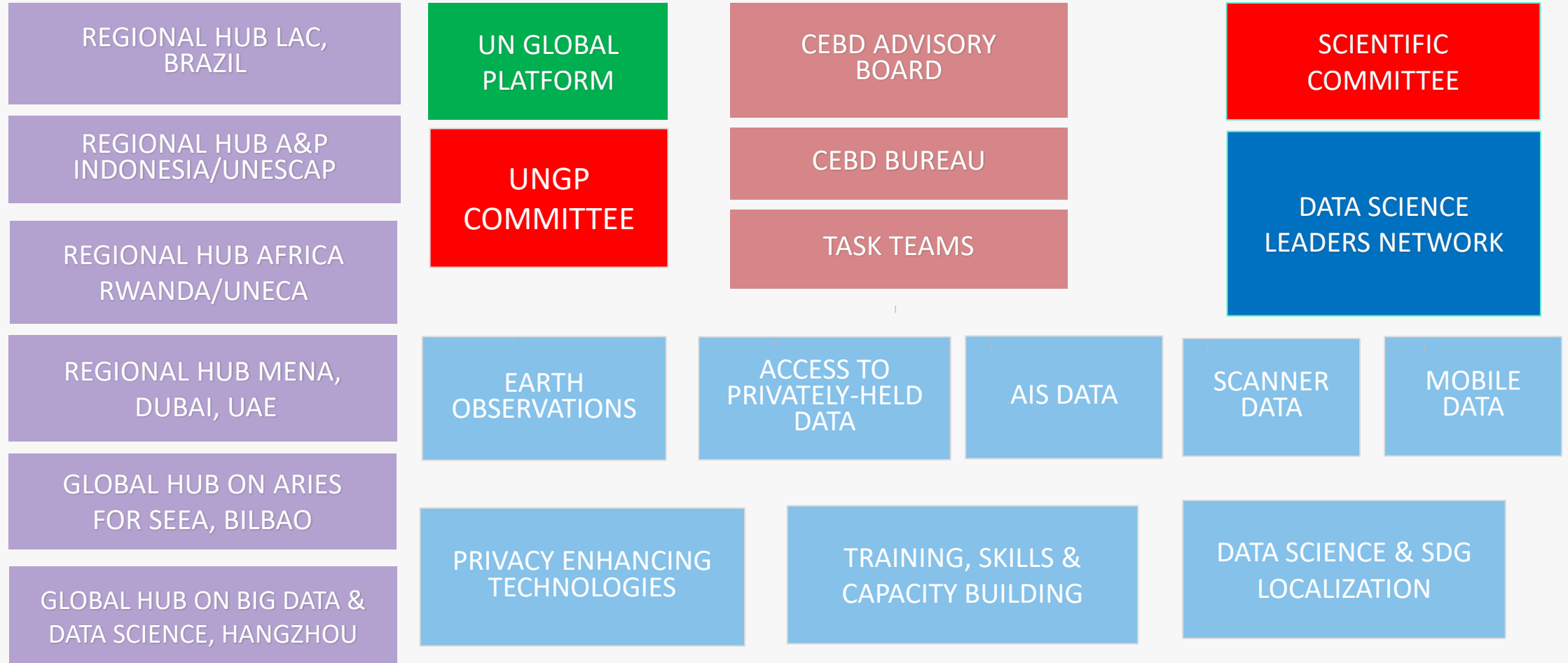


UN Committee of Experts on Big Data and Data Science for official statistics

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





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

Supports Services

Cloudflare

Net App

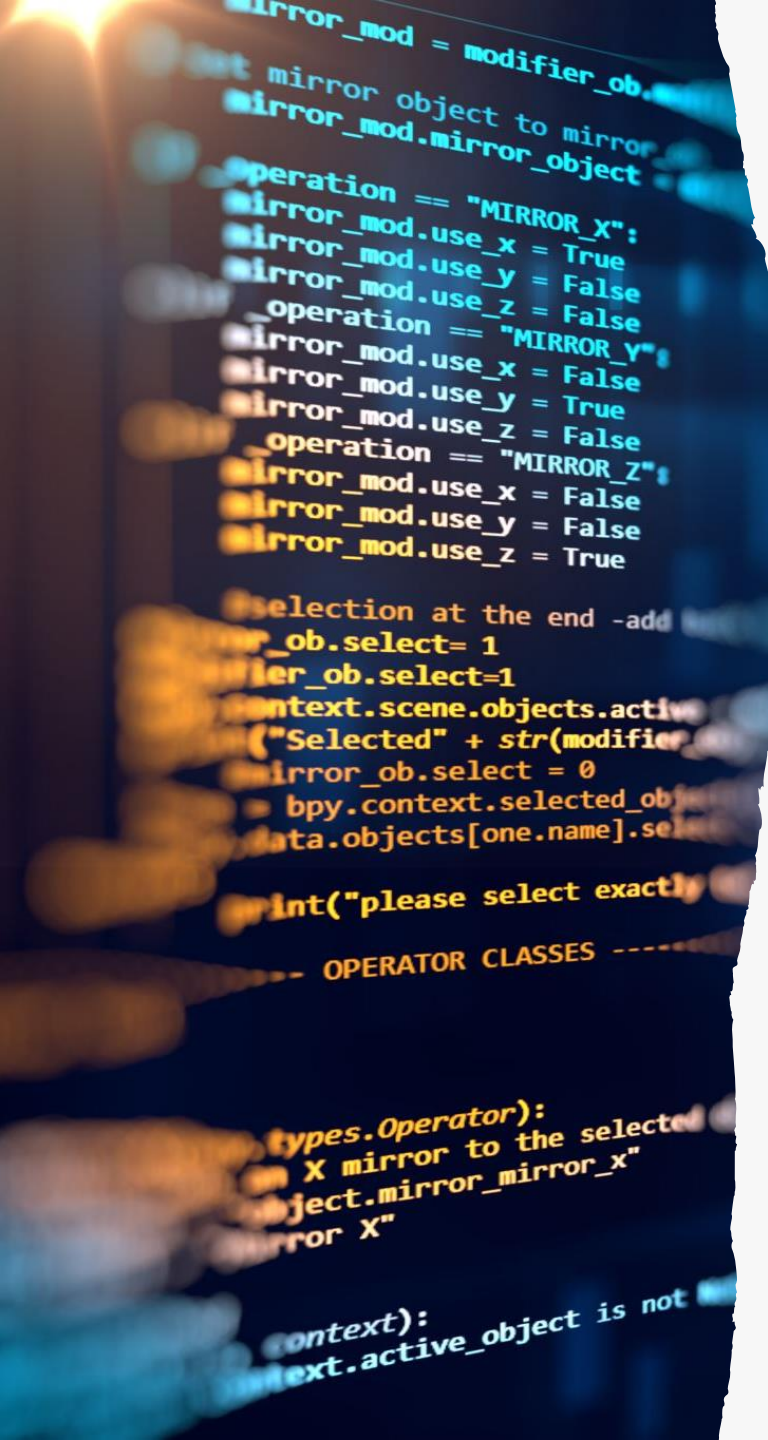
Google suite

Developer Services

Jupyter Notebook

R, Python

Colab



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

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

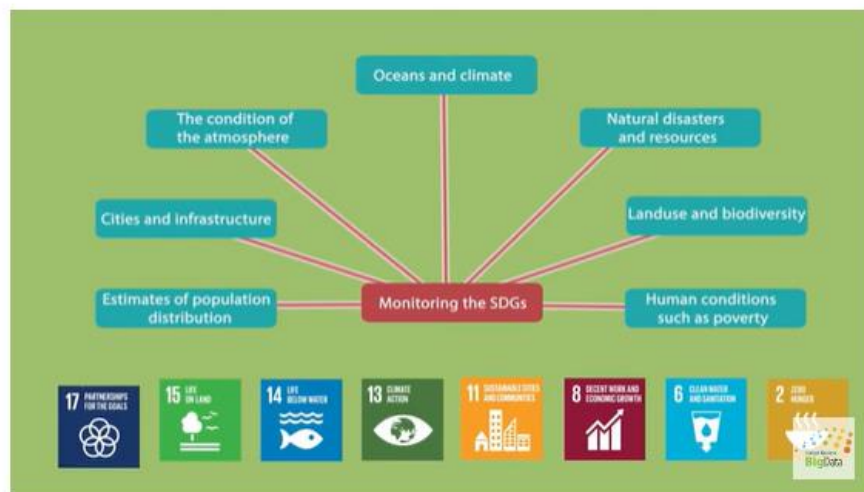
Environment

Agriculture

SDG indicators

Privacy protection

Task Team on Use of Earth Observations for Official Statistics



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

Task Team Report

[Satellite Imagery and Geospatial Data Task Team report](#)

Use Cases

- [Digital Earth Africa](#)
- [Use of EO data in Agriculture Statistics of Statistics Canada](#)

More information

References

- [Report of the Global Working Group on Big Data for Official Statistics](#)
- [Results of the UNSD/UNECE Survey on organizational context and individual projects of Big Data](#)
- [Big data and modernization of statistical systems](#)

Publications



UNITED NATIONS

United Nations Statistical Commission

**United Nations Committee of Experts on Big Data and Data Science for
Official Statistics (UN-CEBD)**

Earth Observation Joint Task Team on Agricultural Production Statistics

Research Sub Task Team

Trusted methods: Lessons Learned and Recommendations from Select Earth Observation
Applications on Agriculture

**Earth Observations for Official
Statistics**

**Satellite Imagery and Geospatial Data
Task Team report**

5th December 2017

2022

[sits: Satellite Image Time Series Analysis on Earth Observation Data Cubes](#)

- Table of contents
- [Preface](#)
- [Setup](#)
- [Acknowledgements](#)
- [Introduction](#)
- [Earth observation data cubes](#)**
- [Operations on data cubes](#)
- [Working with time series](#)
- [Improving the quality of training samples](#)
- [Machine learning for data cubes](#)
- [Classification of raster data cubes](#)
- [Bayesian smoothing for post-processing](#)
- [Validation and accuracy measurements](#)
- [Uncertainty and active learning](#)
- [Ensemble prediction from multiple models](#)
- [Object-based time series image analysis](#)
- [Technical annex](#)
- [References](#)

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 files, 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.

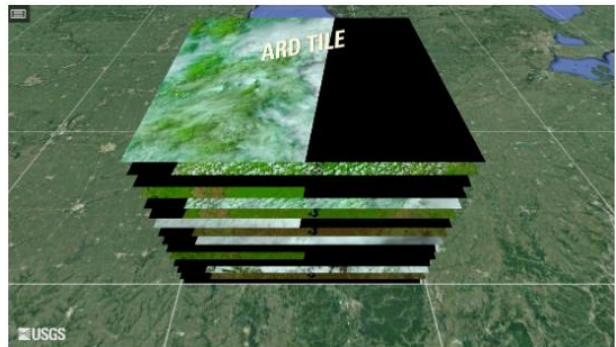


Figure 9: ARD image collection (Source: USGS. Reproduction based on fair use doctrine).

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 $110 \times 110 \text{ km}^2$ with a 10 km

On this page

- [Earth observation data cubes](#)
- [Analysis-ready data image collections](#)**
- [ARD image collections handled by sits](#)
- [Regular image data cubes](#)
- [Creating data cubes](#)
- [Assessing Amazon Web Services](#)
- [Assessing Microsoft's Planetary Computer](#)
- [Assessing Digital Earth Africa](#)
- [Assessing the Brazil Data Cube](#)
- [Accessing Harmonized Landsat-Sentinel collections](#)
- [Defining a data cube using ARD local files](#)
- [Defining a data cube using classified images](#)
- [Regularizing data cubes](#)

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

Methodological Guides on the use of Mobile Phone Data (2022)

- [Displacement and Disaster Statistics](#)
- [Dynamic Population Mapping](#)
- [Measuring the Information Society](#)
- [Migration Statistics](#)
- [Tourism Statistics](#)

Publications

Handbook on the use of Mobile Phone data for Official Statistics (2019)

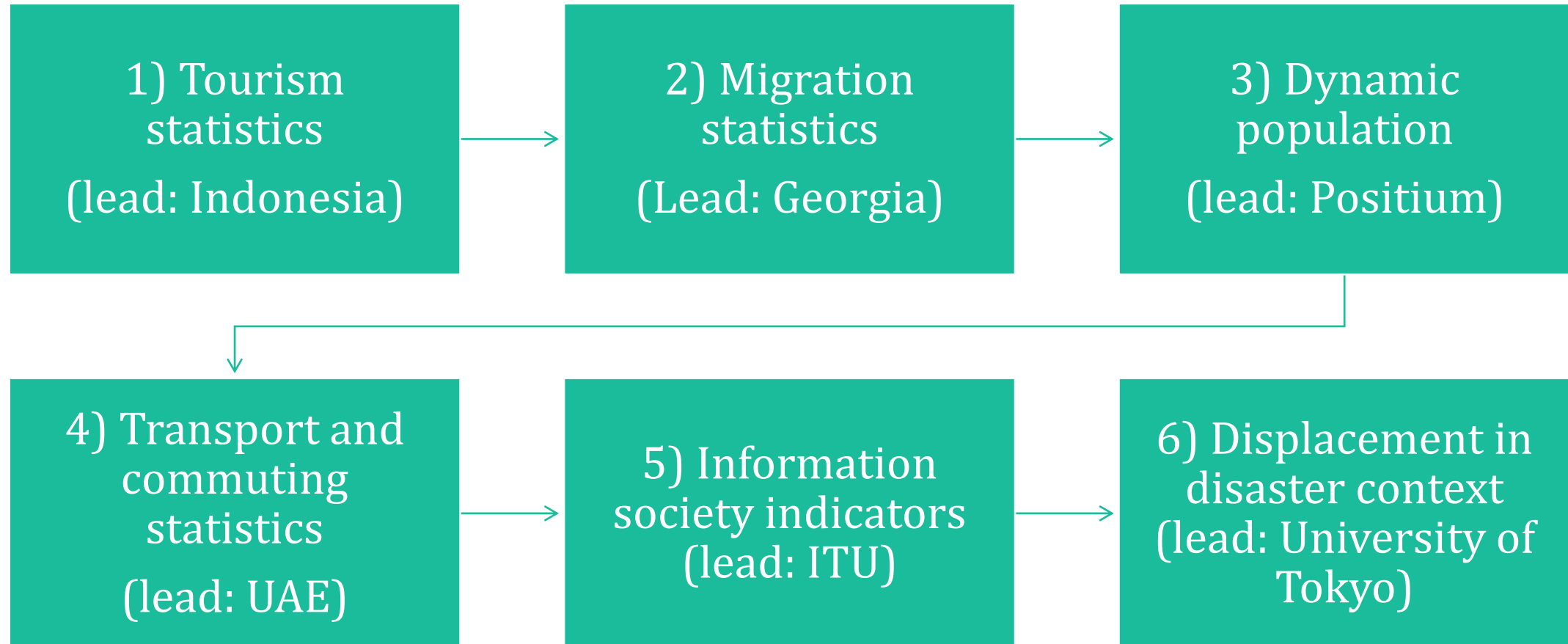
- [English](#)
- [Russian](#)

Events

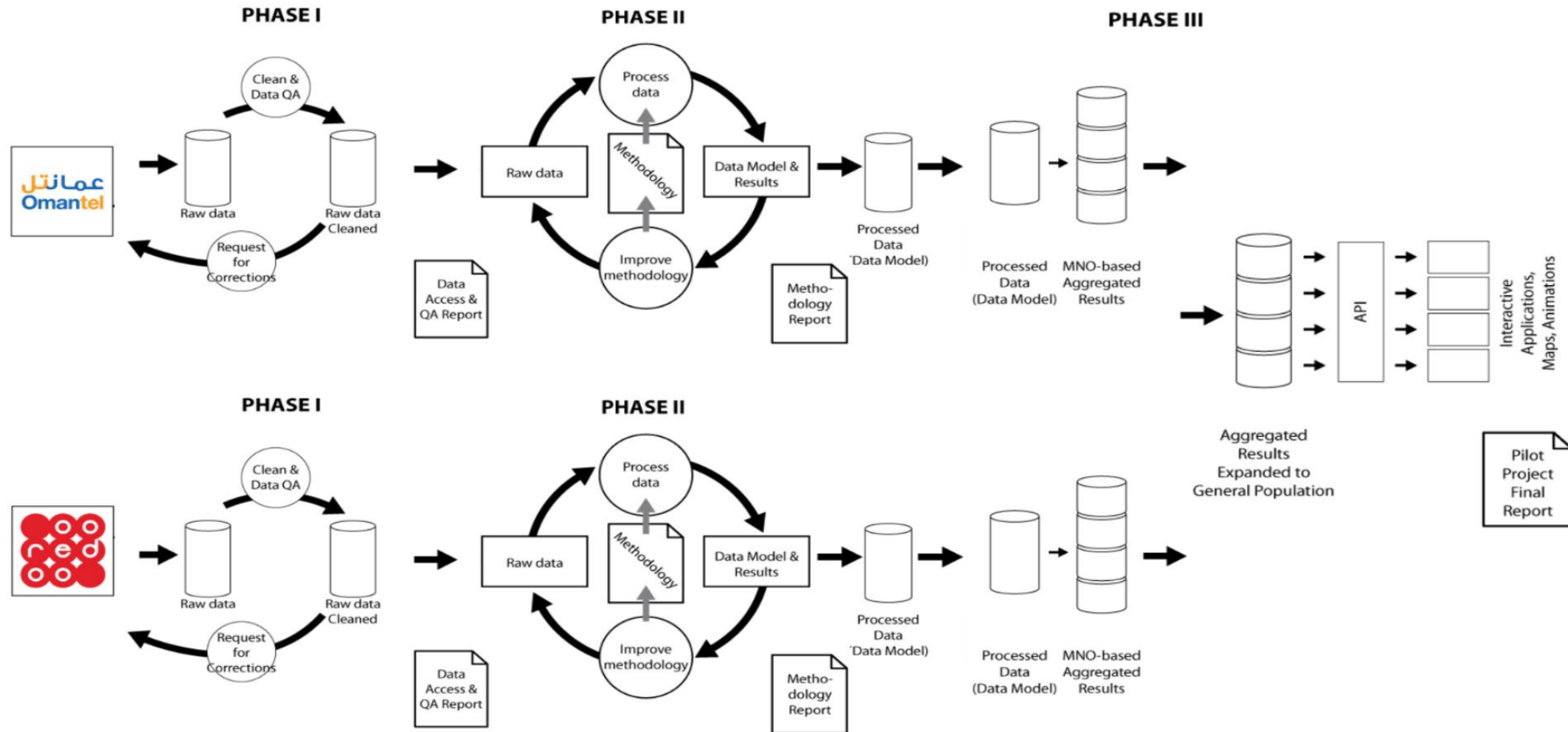
[MPD Session at the 7th International Conference on Big Data](#)

📍 Yogyakarta, Indonesia 📅 8 Nov 2022

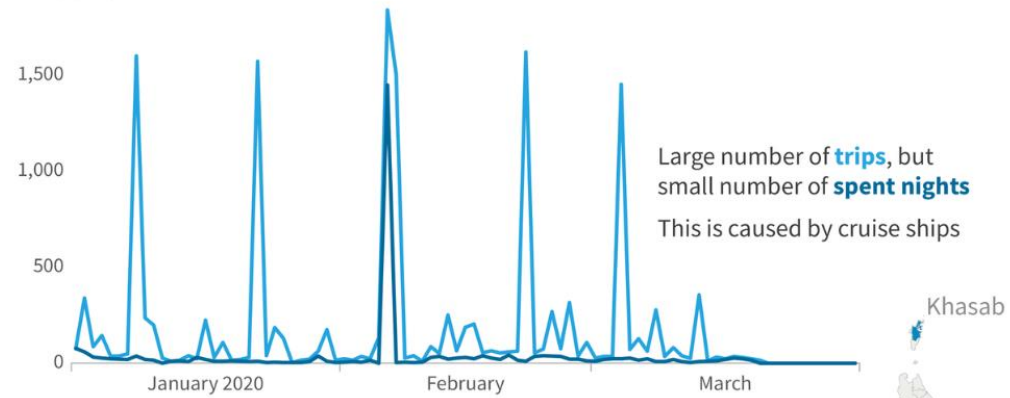
Handbooks



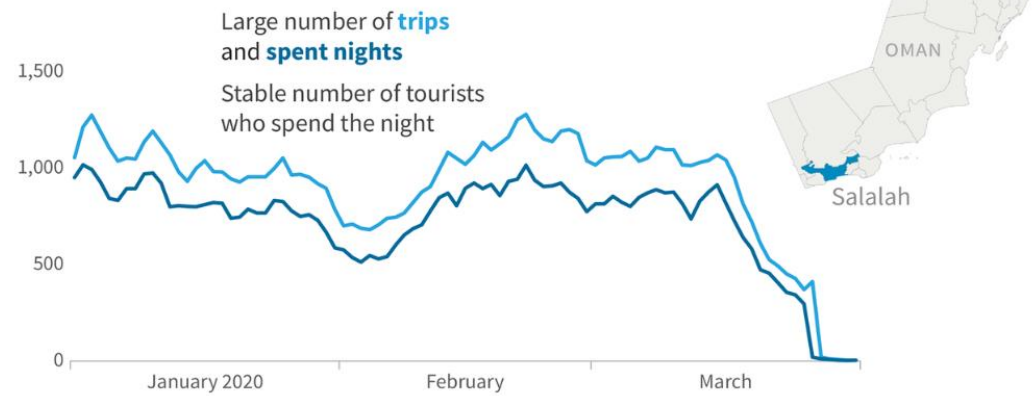
Data Flow



GERMAN TOURISTS in Khasab



Salalah



Task Team on Use of Scanner Data and Webscraping for Price Statistics

The screenshot shows a web browser displaying a page on the UN Statistics Wiki. The page title is "Data filtering and missing prices" under the "UN-CEBD - Scanner Data Wiki" space. The page content includes an overview of data filtering methods, a list of section contents, and a table of page information. The left sidebar shows a navigation menu with various categories like "Pages", "Blog", "Calendars", and "PAGE TREE".

Pages / UN-CEBD Task Team on Scanner Data / Handbook on utilising new data sources in the production of consumer price statistics 311 views

Data filtering and missing prices

Created by Clarence Lio (UNSD), last modified by Tanya Flower yesterday at 6:24 PM

This section covers some common methods used to filter the data before it is used to calculate indices, as well as a summary of how missing prices should be treated.

Section contents:

- **Outlier filter** — Outlier filters aim at excluding or correcting extreme price increases or price decreases, typically compared to the previous period, from the price index calculation
- **Dumping filter** — Dumping filters aim at eliminating the downward pressure of clearance prices on the index.
- **Low sales filter** — The low sales filter ensures that products with small expenditure shares do not unduly influence the index results.
- **Current practice of data filtering** — This page contains a summary of how NSOs currently approach data filtering.
- **Treatment of missing prices** — This page contains information on how to treat missing prices through product-level imputation methods.

Page information

Version	v1
Last updated on	2024-03-01
Summary of changes	
Page History	

Overview

The scale of these new data sources mean that new methods are required to automatically identify and remove erroneous observations, as the traditional manual scrutiny of flagged observations is not feasible. Traditional data filtering often focuses on removing observations that are thought to be the result of erroneous data collection or that have a large impact due to small sample sizes.

New data sources tend to be much larger than conventionally collected surveys and reduce standard error due to sample size. Given the nature of scanner data, it is also less prone to ad hoc data collection errors compared with web scraped data, although some errors still can exist. Filters for non-survey data may be useful for addressing these issues with the dataset (for example, decimal point issues) and ensuring the reliability of the index results.

When processing new data sources for consumer price indices, three types of data filters that are applied directly to the product level data have been established as best practice:

- outlier filter
- dumping filter
- low sales filter

The outlier filter focuses only on extreme price increases or decreases whereas the dumping filter and low sales filter additionally take the development of sales quantities into account. The settings of these data filters depend not only on the index formula that will be used but also on the type of data these filters are applied to. Typically, these data filters are used directly before the index calculation. Thereby, the order of filtering may have an impact on the index results. Usually, the outlier filter is applied first and then the dumping or low sales filter. For more detail on how these filters relate to the choice of index method and data source, please see the specific pages linked above.

For products identified as outliers it may be appropriate to consider *imputation*. For those products filtered out by the dumping or low sales filter, these should not be imputed.

Since web scraping data comprise of information only on prices but not sales quantities, the dumping filter and low sales filter cannot be applied in the typical manner. However, there is currently ongoing research into methods to *estimate the expenditure shares of web scraped articles*.

Over time, it is important to monitor whether the applied data filters are set appropriately. For example, it should be checked periodically if the share of filtered products is too high or if some certain product groups are filtered out systematically and unjustifiably. In this case, the data filters should be adjusted or there should be an option implemented into the production system, which allows the data filter to be overruled.

Data filtering is an area of active research and it is likely that different data sources and/or suppliers require different approaches to data filtering based on their own existing quality controls and circumstances. As such, National Statistical Offices (NSOs) are recommended to conduct their own empirical research using these approaches before deciding if any data filtering are required and if so which approach is most suitable.

Like Be the first to like this

No labels

Write a comment...

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

This is a beta version that has been made available for public use and comment. We welcome your feedback at IMF-PortWatch@IMF.org

IMF | PORTWATCH
A PARTNERSHIP WITH
OXFORD UNIVERSITY

[ABOUT US](#) | [SIGN UP FOR ALERTS](#) | [ACCESS DATA](#) | [IMF](#)

[PORT MONITOR](#)

[RECENT DISRUPTIONS](#)

[SPILLOVER SIMULATOR](#)

[CLIMATE SCENARIOS](#)

DISRUPTION MONITOR

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 Strait) due to attacks on commercial ships

Event date: Reduced 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 systemically important shipping lane that facilitates about 15 percent of global maritime trade volume and over 22,000 transit calls annually. Further details are discussed in [this IMF blog post](#).

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 from Russia to Asia.

Chokepoints in affected area: Bab el-Mandeb Strait; Suez Canal; Cape of Good Hope

Suez Canal (located at the northern entrance to the Red Sea)

Transit Calls | Transit Trade Volume

Suez Canal: Daily Transit Calls

[Export Data](#)

Zoom 1m 3m 6m YTD 1y All

21 Mar 2024 → 17 Apr 2024

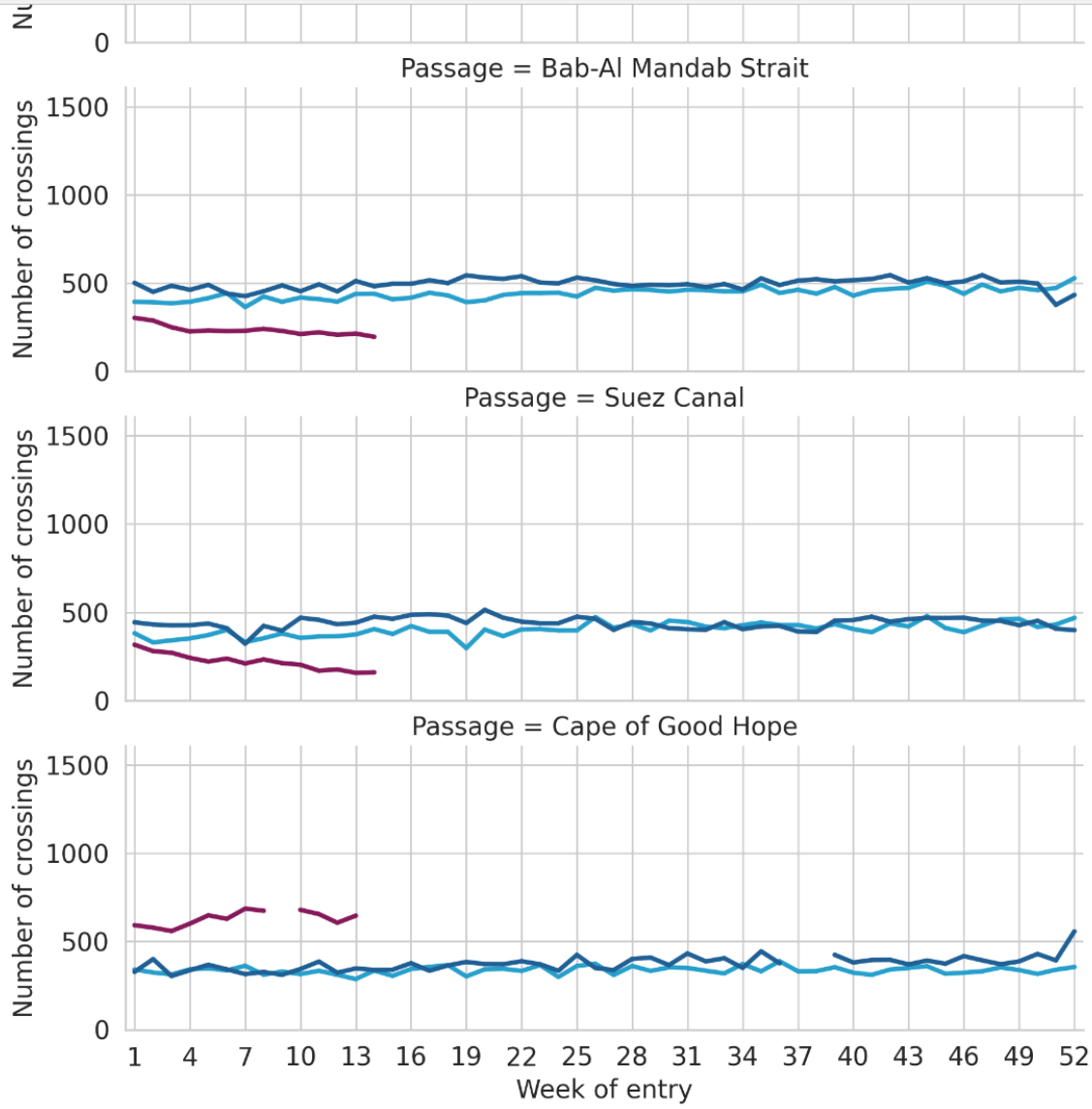
Ship Traffic in Critical Maritime Passages

[Data Science Campus](#) [Department for Business and Trade](#)

April 24, 2024

Categories: [Data Science Campus](#), [Emerging Issues](#), [International](#), [Trade](#), [Shipping and Global Supply Chains](#)





Task Team on Privacy Enhancing Technologies for Official Statistics

Privacy-Enhancing Technologies

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



HOME ABOUT ▾ EVENTS TASK TEAMS ▾ REGIONAL HUBS ▾ UN GLOBAL PLATFORM

A banner image for the 'Privacy-Enhancing Technologies' page. It features a large, metallic padlock in the foreground, with tree branches and leaves in the background. The text 'Privacy-Enhancing Technologies' is prominently displayed in white, with the subtitle 'Task Team of the UN Committee of Experts on Big Data and Data Science for Official Statistics' below it. At the bottom left, there is a breadcrumb trail: 'Home > Task Teams > Privacy-Enhancing Technologies'.

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 technique.

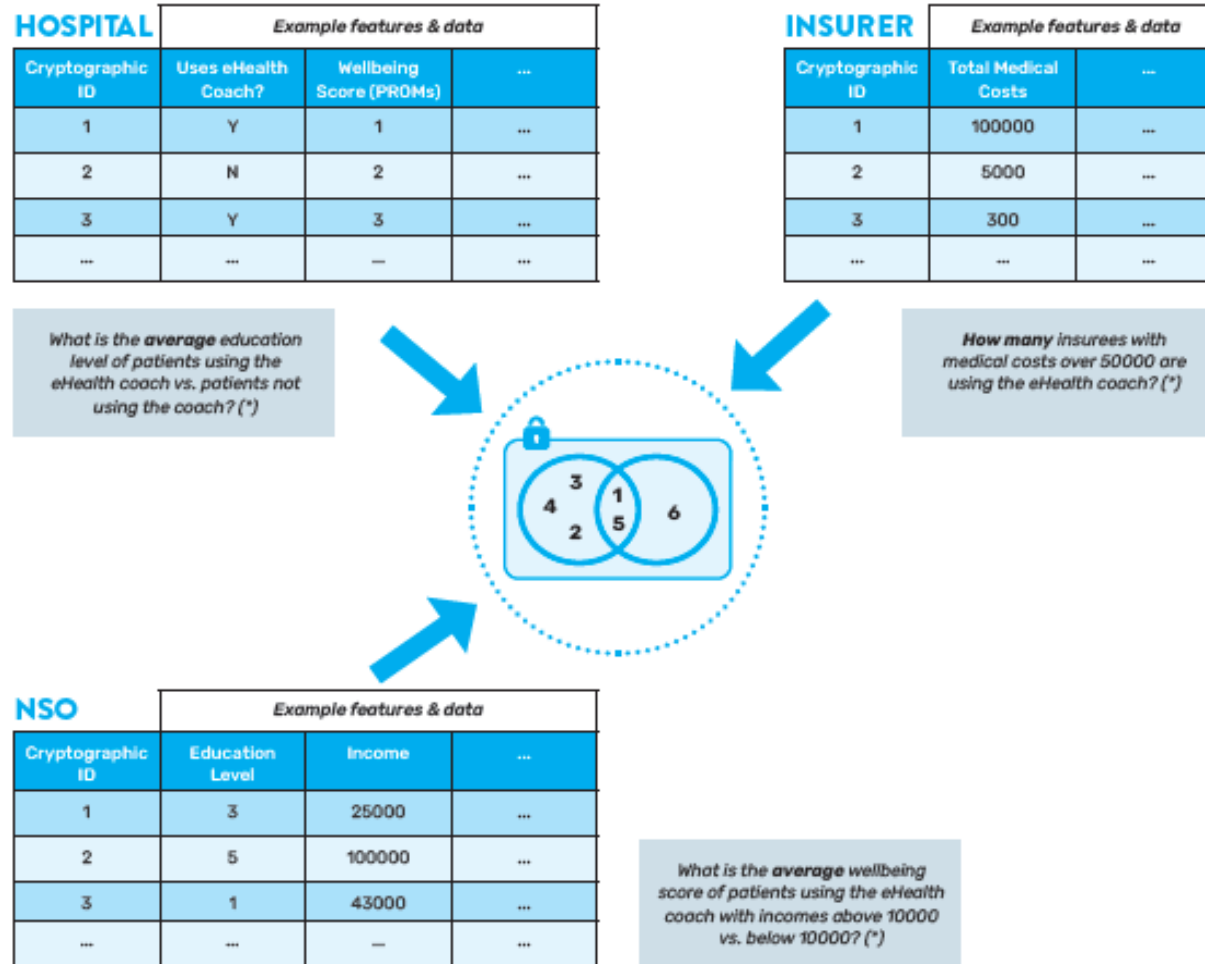
The task team has been active since April 2018 and has released the UN Privacy Preserving Techniques Handbook. This document describes motivations for privacy-preserving approaches for the statistical analysis of sensitive data; presents examples of use cases where such methods may apply; and describes relevant technical capabilities to assure privacy preservation while still allowing analysis of sensitive data. It summarizes current best practices that ensure the protection and sharing of sensitive information. Going

Publications



CASE STUDY DESCRIPTION

HIGH LEVEL FUNCTIONAL PERSPECTIVE



*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



Data Science Campus | April 4, 2024

Categories: Data Science Campus, Projects, Synthetic data and PETs



The PPRL toolkit demonstrates a layered approach to security, which has been called the 'Swiss cheese' model

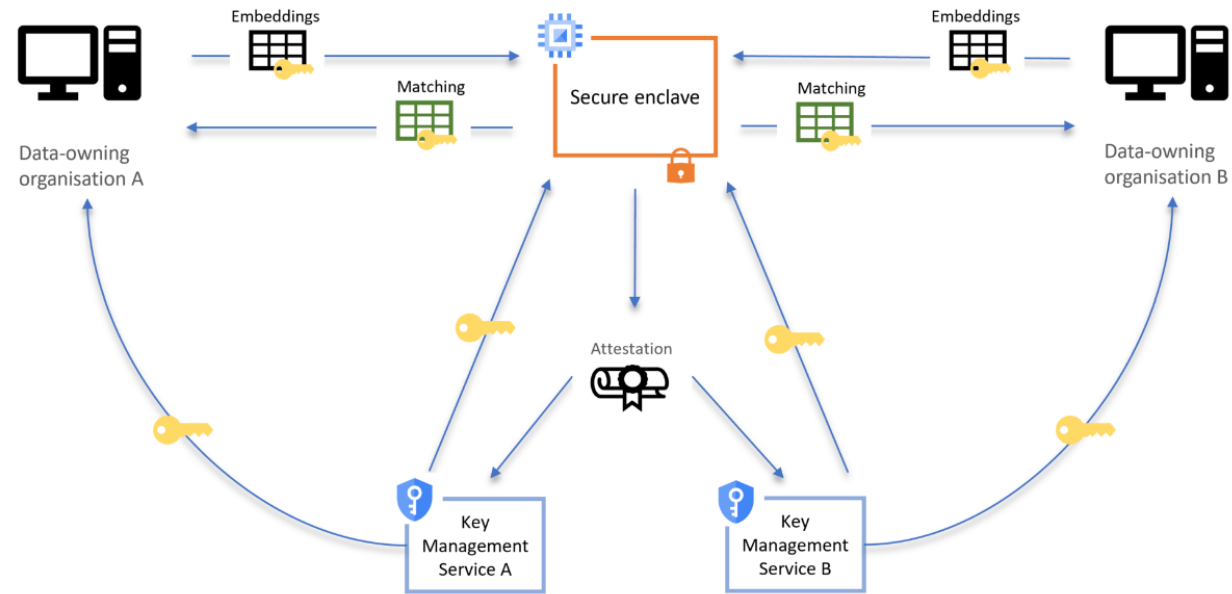
Overview

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

Figure 1: Example diagram for the PPRL server architecture



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

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

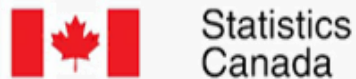
- **Automation** of the statistical production processes (increase efficiency and improve quality)
- **Supplementary indicators** produced for emerging issues to provide additional insights
- **Changing statistical production:**
Example – webscraping of prices from the internet combined with traditional price surveys to produce regular consumer price indices

Reproducible analytical pipelines

Supplementary
analysis and
insights

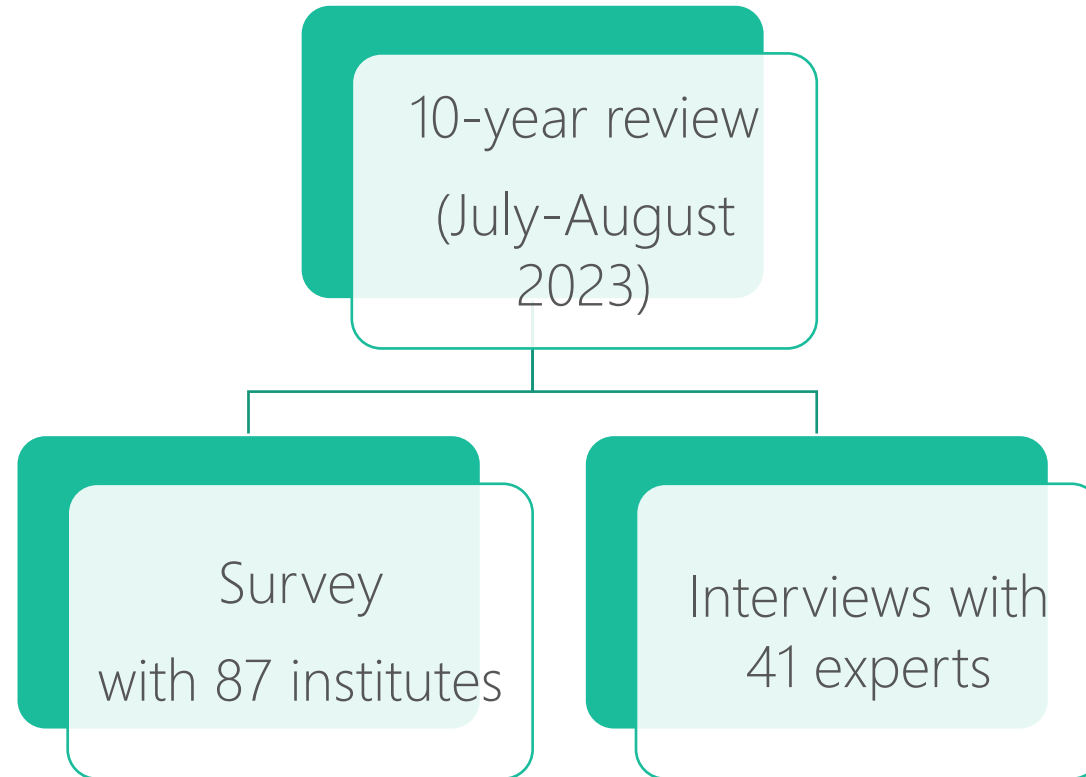
Transformation
of Statistical
Production
Process

UNCEBD partners with about 100 institutes



Survey : 10-year review on use of Big Data

Big Data and Data Science for Official Statistics



Geographical breakdown of institutes

Africa	14
Asia	21
Europe	30
Latin America and the Caribbean	9
Middle East	7
North America	2
Oceania	4

Questions for Survey and Interview

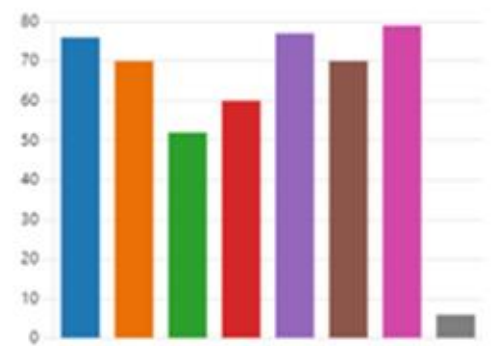
GAMSO structure of the **Survey** (Generic activity model of statistical organizations)

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

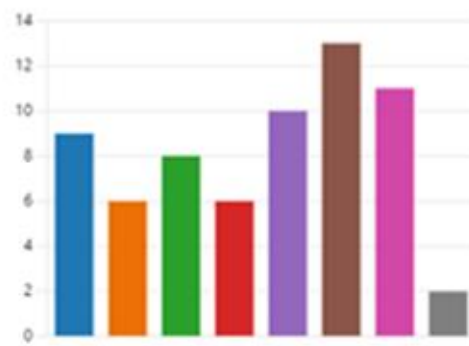
Categories of the **interview** **questions**

- 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

Strategic Vision



National Statistical Offices



International Organizations

- Access to data from private sector
- Modernization of statistical legislation
- Use of Satellite data
- IT Cloud computing and services
- Data privacy protection
- Capacity development in data science or data engineering
- Collaboration with universities or private partnerships
- Other

10. High-level goals of multi-year plan

Please check all those topics below which are part of your innovation strategy

IT Management



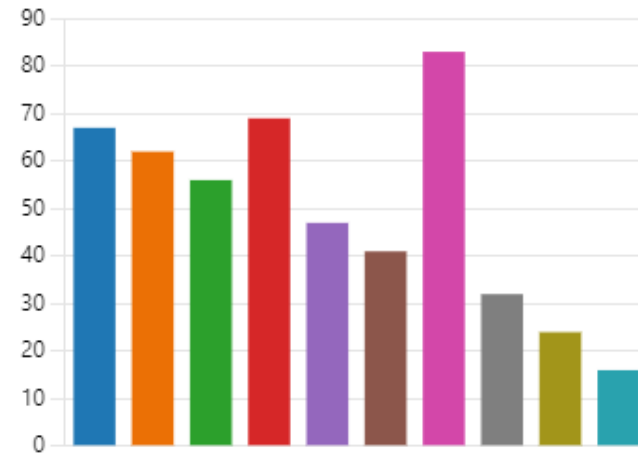
38. Infrastructure

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

● Satellite data	67
● Mobile phone data	62
● Retail store scanner data	56
● Webscraping data	69
● Credit card / Payment card data	47
● Citizen generated data / Citizen ...	41
● Additional administrative data (...)	83
● Smart meter data	32
● Other private sector data, pleas...	24
● Other data sources, please speci...	16

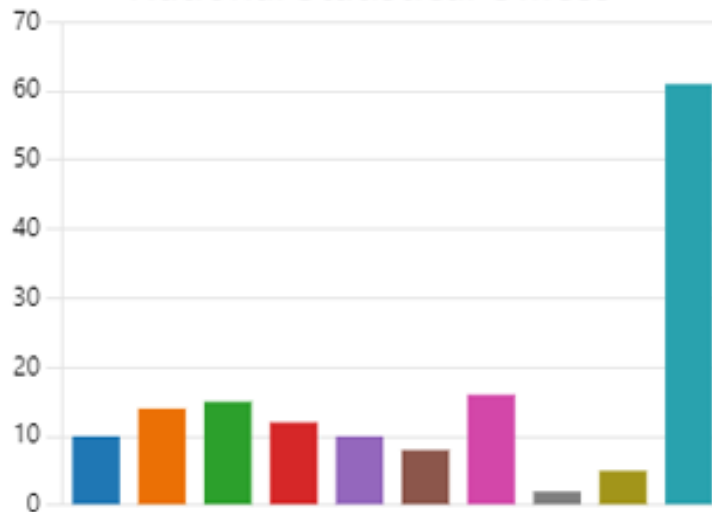


20. Alternative data sources, including Big Data or webscraping and other privately held data

What other data sources do you use or are you considering using in the future?

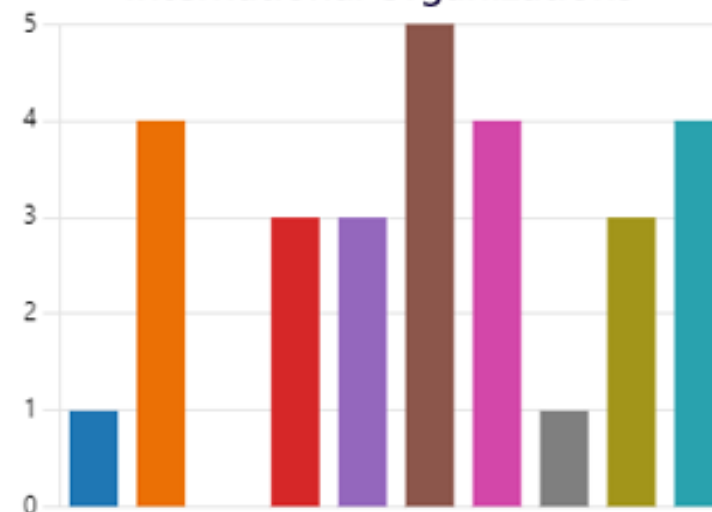
Task team participation – in which task teams has your office been actively participating

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 webscraping 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

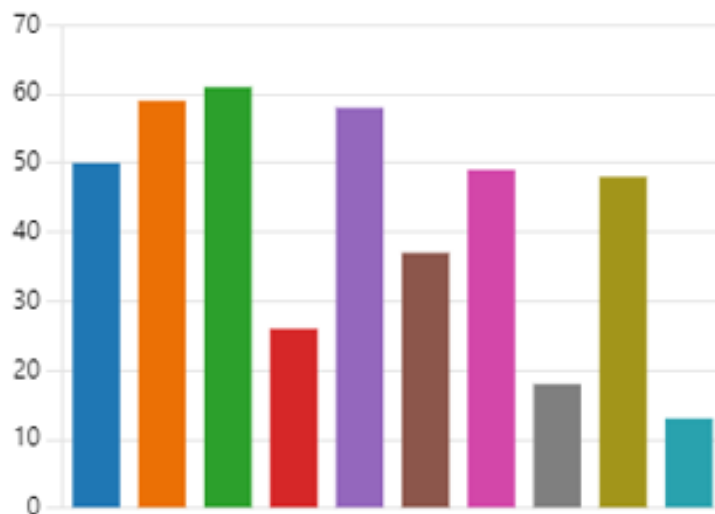
International Organizations



- 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

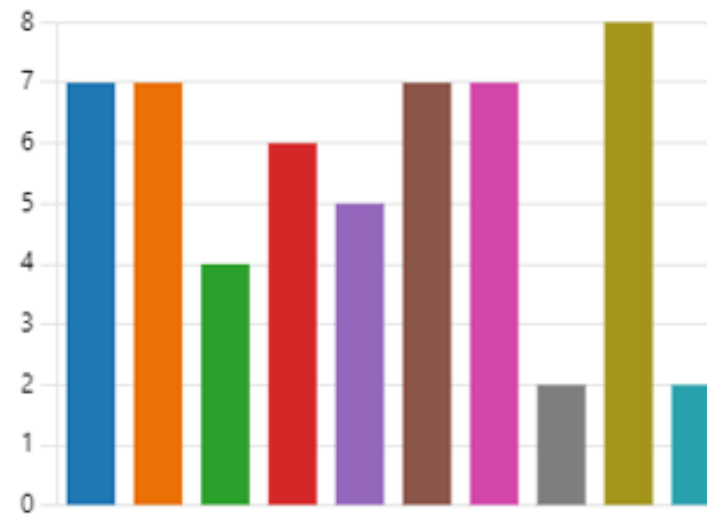
Task teams of interest – in which task teams would your office be interested to participate

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 webscraping 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

Summary of Survey Results



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.

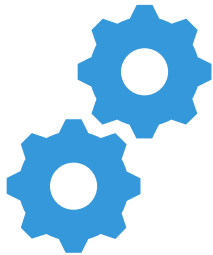


Correspondingly, more than 4 out of 5 NSOs have updated or are in the process of updating their statistical legislation to facilitate access to privately held data.



About half of NSOs and IOs are actively developing data science capabilities in their institutes.

Summary of Survey Results



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.

Recommendation – Communication

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.

Recommendation – Updating the UNCEBD mandate

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**