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

United Nations BigData

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

UNGP COMMITTEE

CEBD ADVISORY BOARD

CEBD BUREAU

TASK TEAMS

SCIENTIFIC COMMITTEE

DATA SCIENCE LEADERS NETWORK

EARTH OBSERVATIONS

ACCESS TO PRIVATELY-HELD DATA

AIS DATA

SCANNER DATA

MOBILE DATA

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

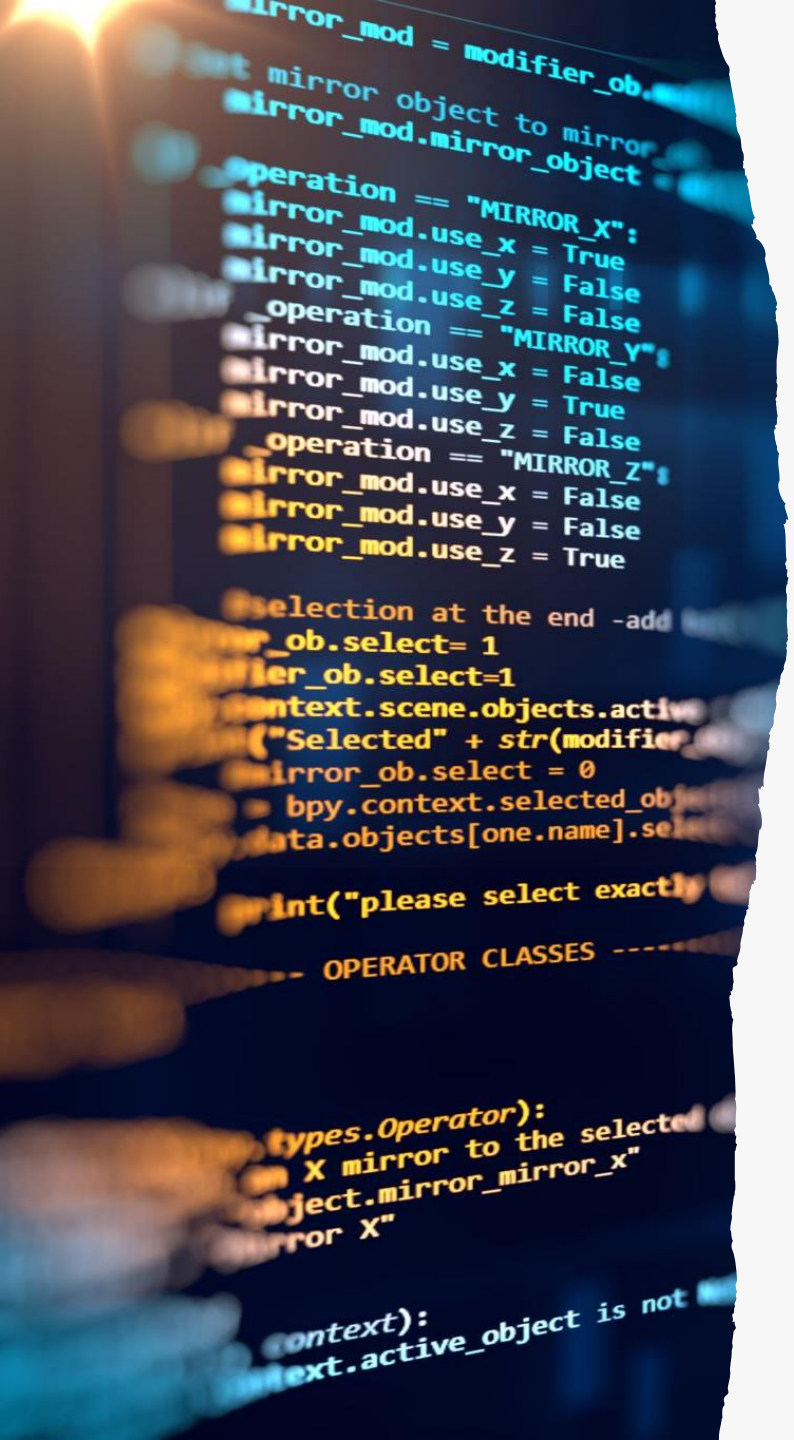
Data as a Service

Trade data

Shipping register data

Port activity

Global group registers



Technology services

Cloud Services

AWS

Google Cloud Platform

MS Azure

Alibaba

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

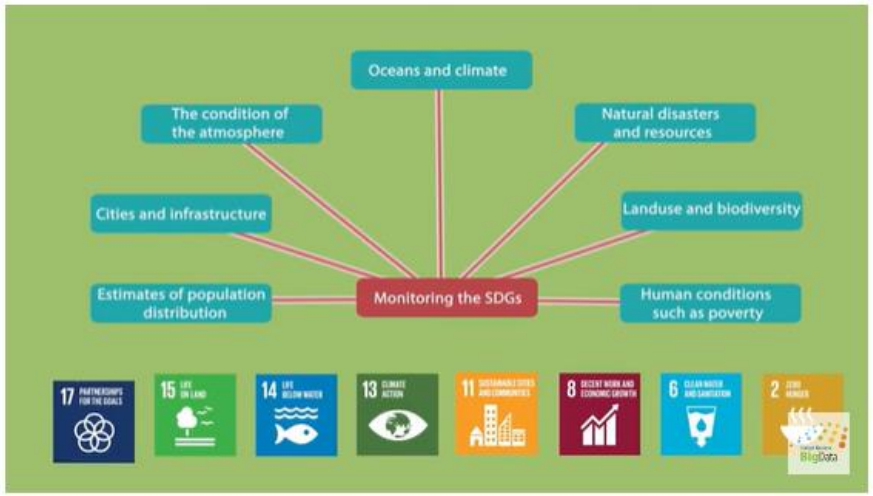
Technology stack

- .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

2022

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 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 multidate 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.

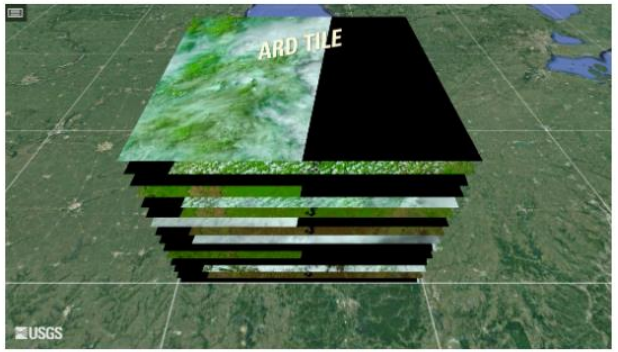


Figure 9: ARD image collection (Source: USGS. Reproduction based on fair use doctrine).

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 110×110 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

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)

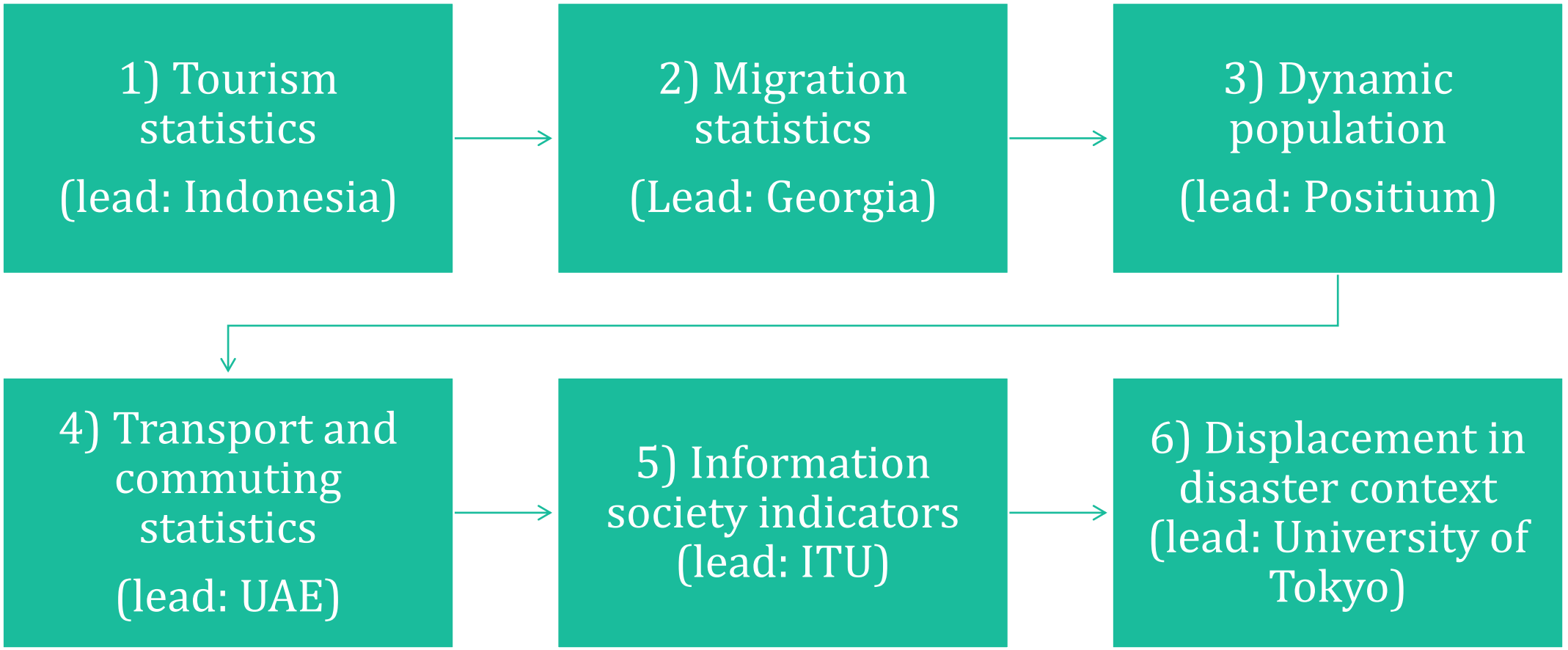
- English
- Russian

Events

MPD Session at the 7th International Conference on Big Data

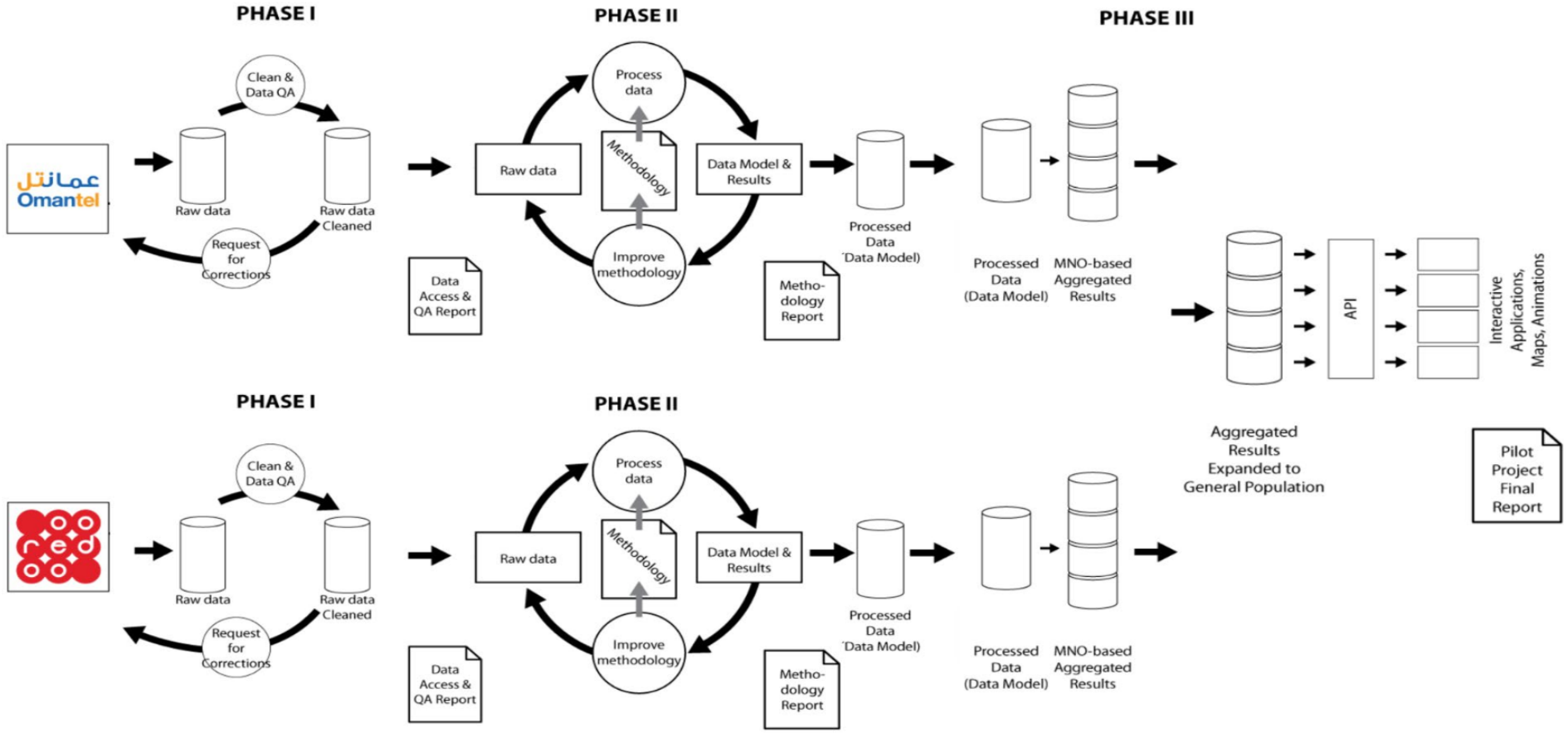
Yogyakarta, Indonesia 8 Nov 2022

Handbooks


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

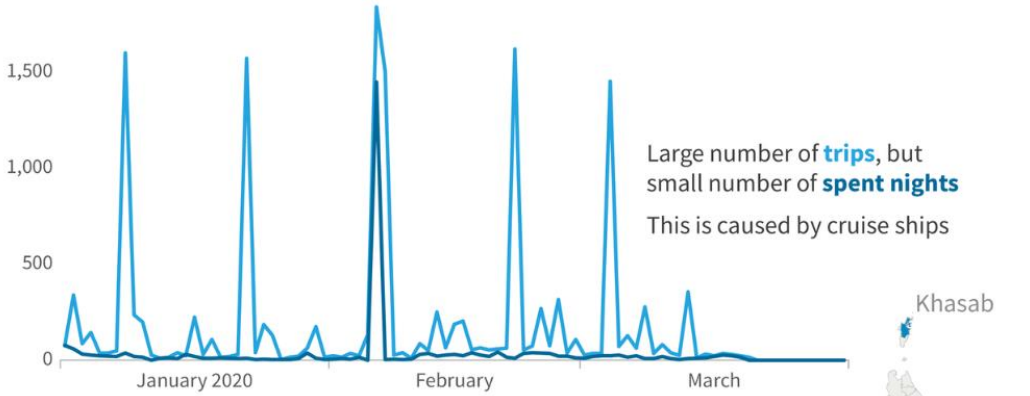
NATIONAL CENTRE FOR STATISTICS & INFORMATION
SULTANATE OF OMAN



GERMAN TOURISTS in

positium

Khasab



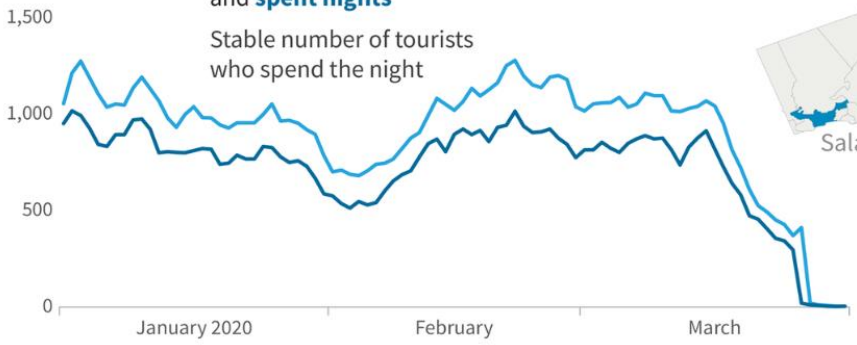
Large number of **trips**, but small number of **spent nights**

This is caused by cruise ships

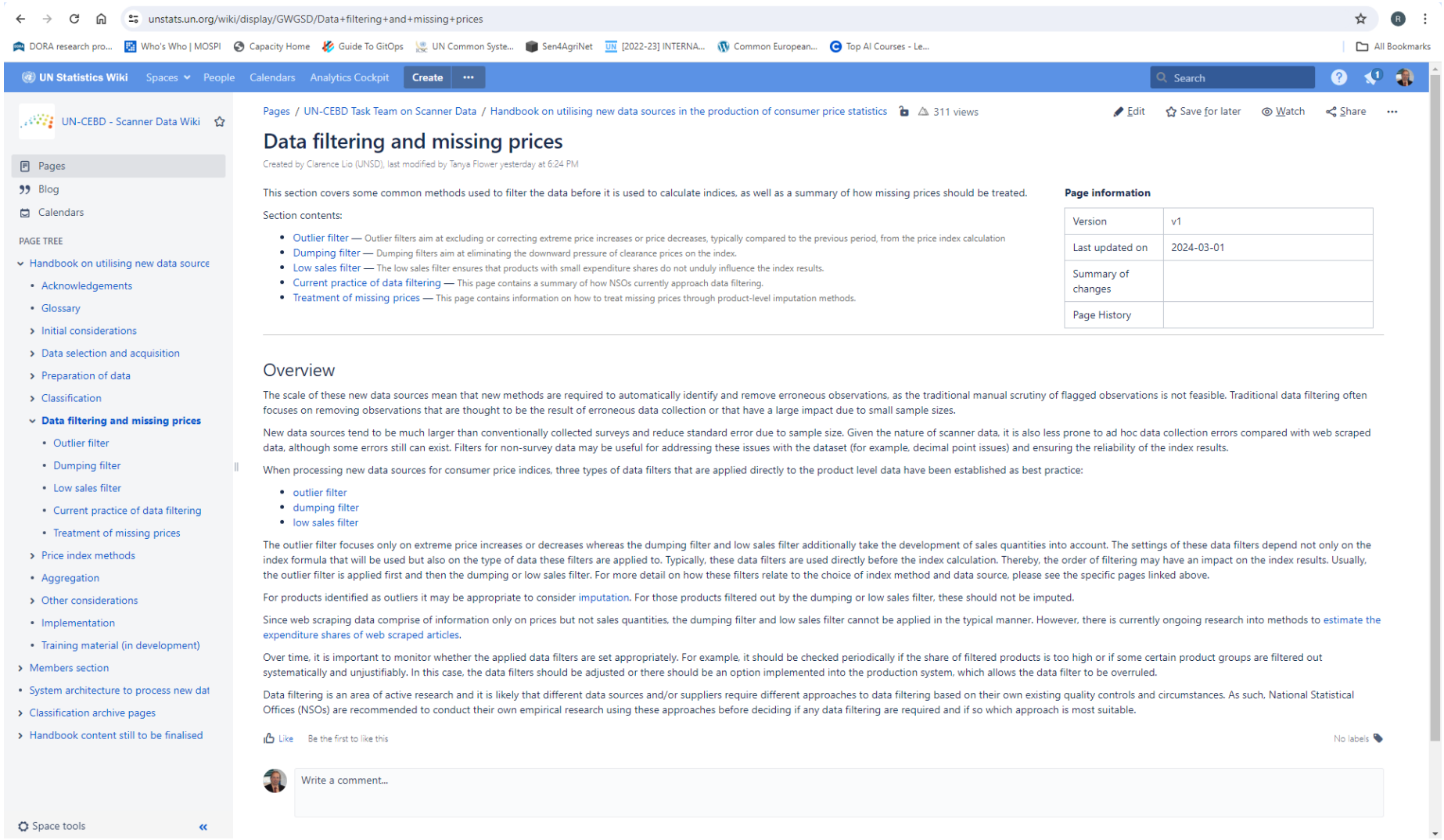
Salalah

Large number of **trips** and **spent nights**

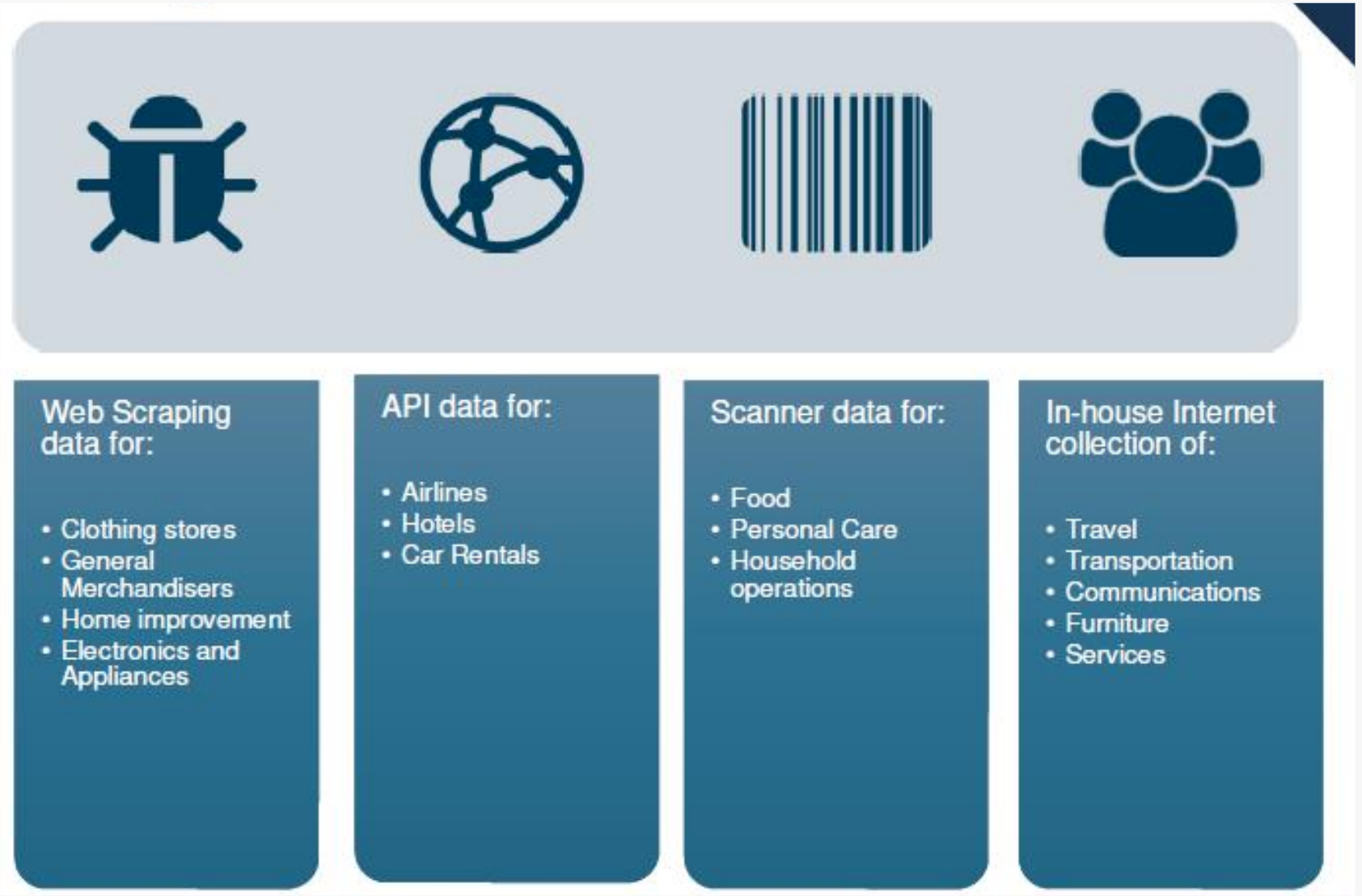
Stable number of tourists who spend the night



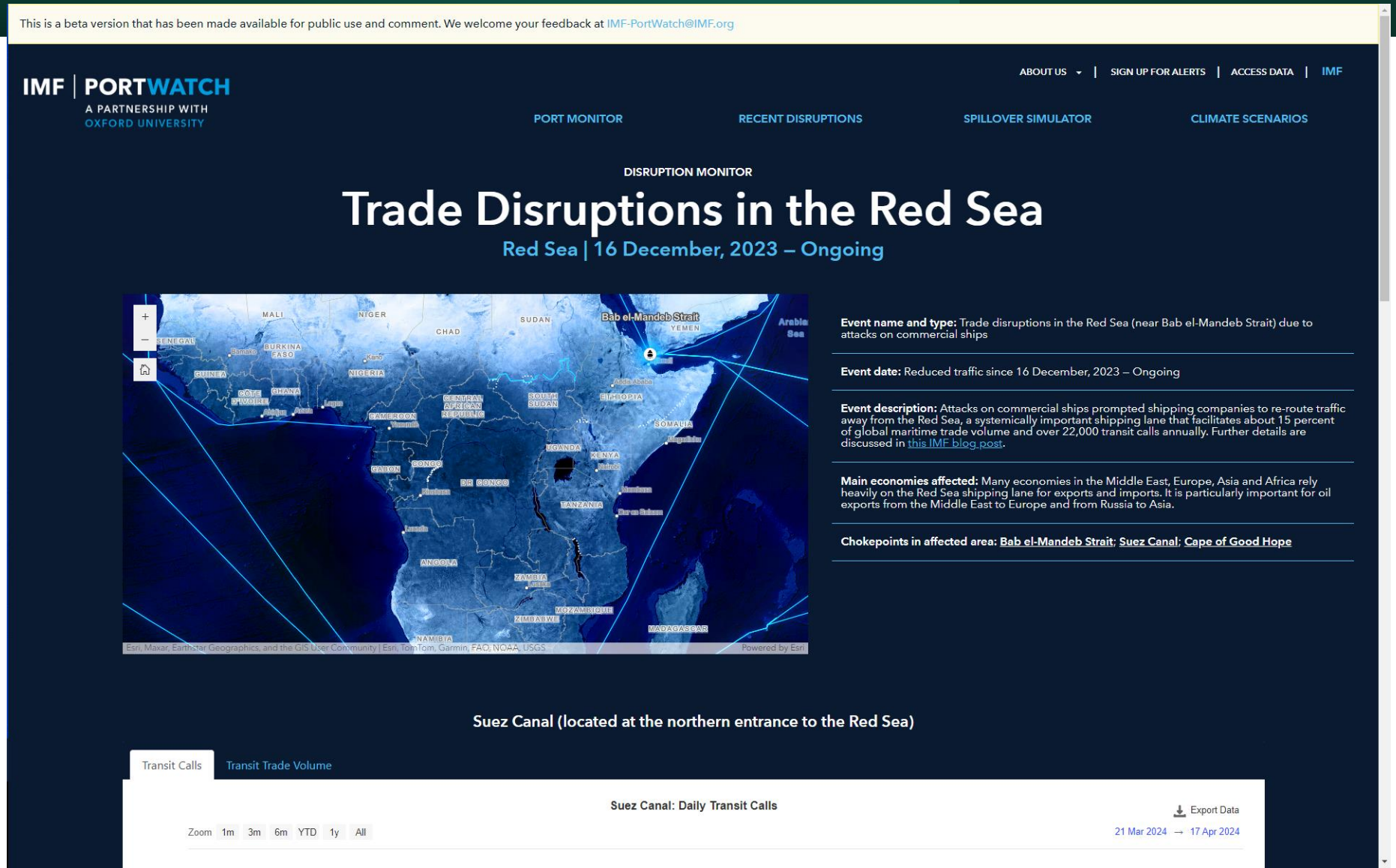
Task Team on Use of Scanner Data and Webscraping for Price Statistics



Prices from digital sources



Task Team on Use of AIS Shipping data for Maritime Trade and Transport Statistics

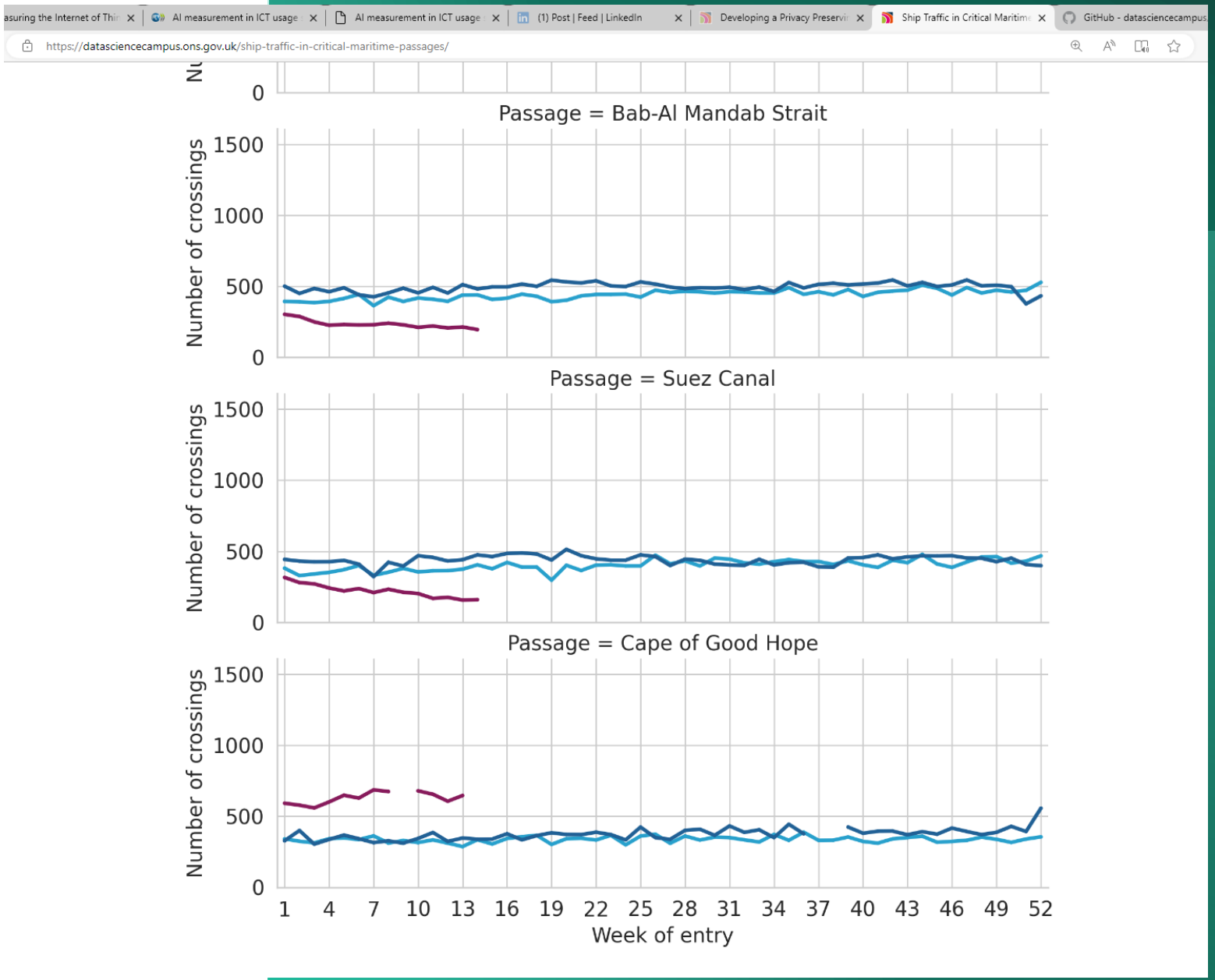


Ship Traffic in Critical Maritime Passages

Data Science Campus Department for Business and Trade

April 24, 2024

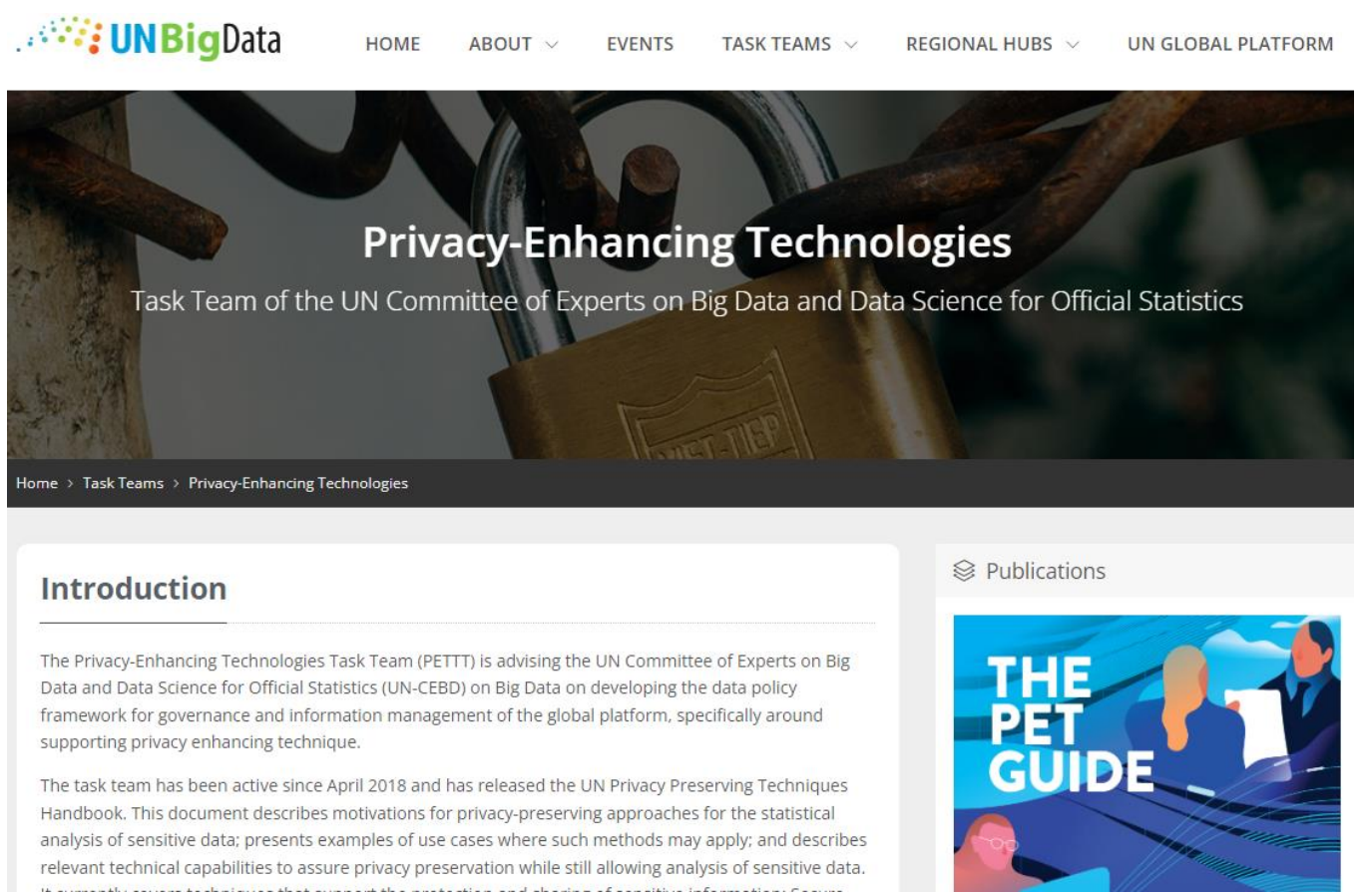
Categories: Data Science Campus, Emerging Issues, International, Trade, Shipping and Global Supply Chains


asuring the Internet of Thi
AI measurement in ICT usage
AI measurement in ICT usage
(1) Post | Feed | LinkedIn
Developing a Privacy Preservi
Ship Traffic in Critical Maritime
GitHub - datasciencecampus
https://datasciencecampus.ons.gov.uk/ship-traffic-in-critical-maritime-passages/
0
Passage = Bab-Al Mandab Strait
Number of crossings
1500
1000
500
0
Passage = Suez Canal
Number of crossings
1500
1000
500
0
Passage = Cape of Good Hope
Number of crossings
1500
1000
500
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52
Week of entry

Task Team on Privacy Enhancing Technologies for Official Statistics

Privacy-Enhancing Technologies

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



CASE STUDY DESCRIPTION

HIGH LEVEL FUNCTIONAL PERSPECTIVE

HOSPITAL

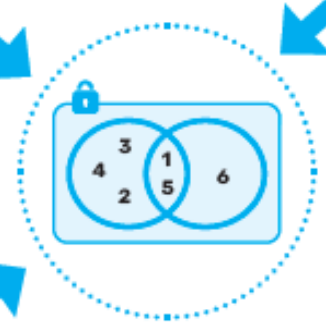
	Example features & data		
Cryptographic ID	Uses eHealth Coach?	Wellbeing Score (PROMs)	...
1	Y	1	...
2	N	2	...
3	Y	3	...
...	...	—	...

INSURER

	Example features & data	
Cryptographic ID	Total Medical Costs	...
1	100000	...
2	5000	...
3	300	...
...	...	...

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

__How many__ insurees with medical costs over 50000 are using the eHealth coach? (*)



NSO

	Example features & data		
Cryptographic ID	Education Level	Income	...
1	3	25000	...
2	5	100000	...
3	1	43000	...
...	...	—	...

*What is the **average** wellbeing score of patients using the eHealth coach with incomes 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

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

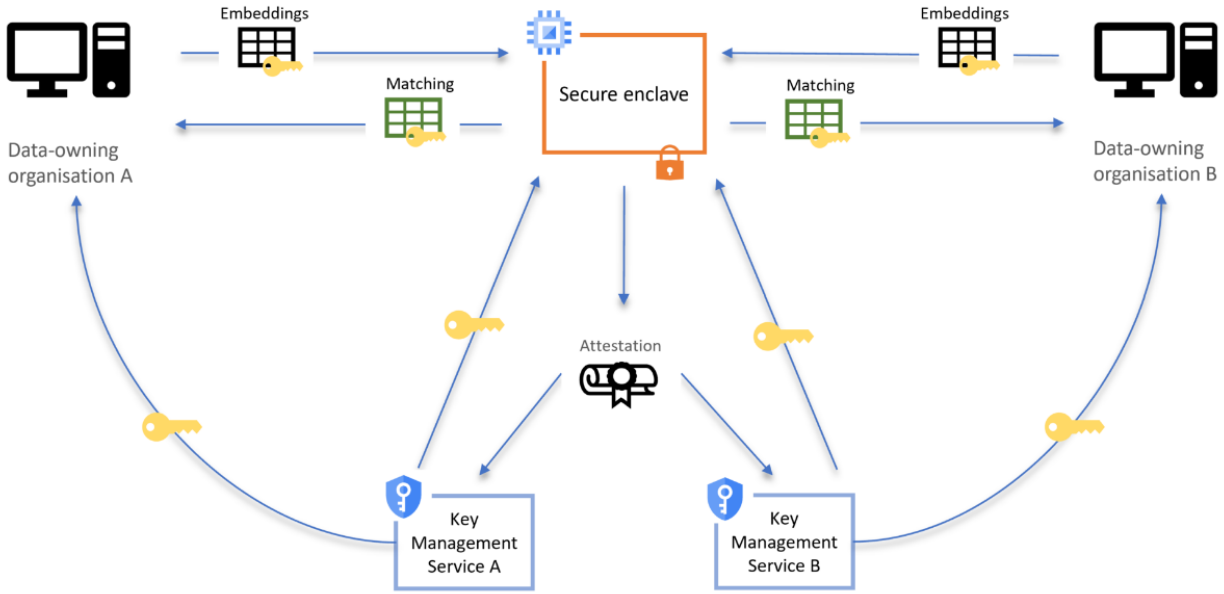
Overview

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

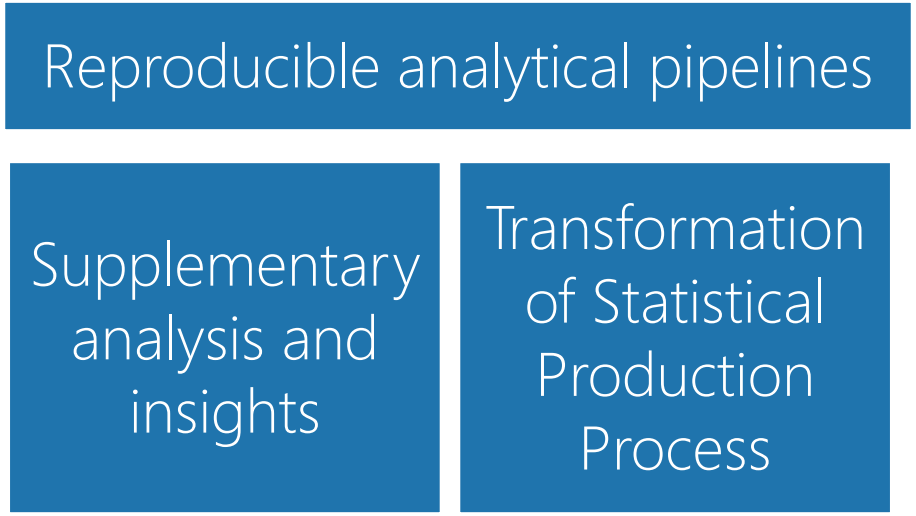
Training in Big Data and Data Science for official statistics

United Nations BigData

- **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

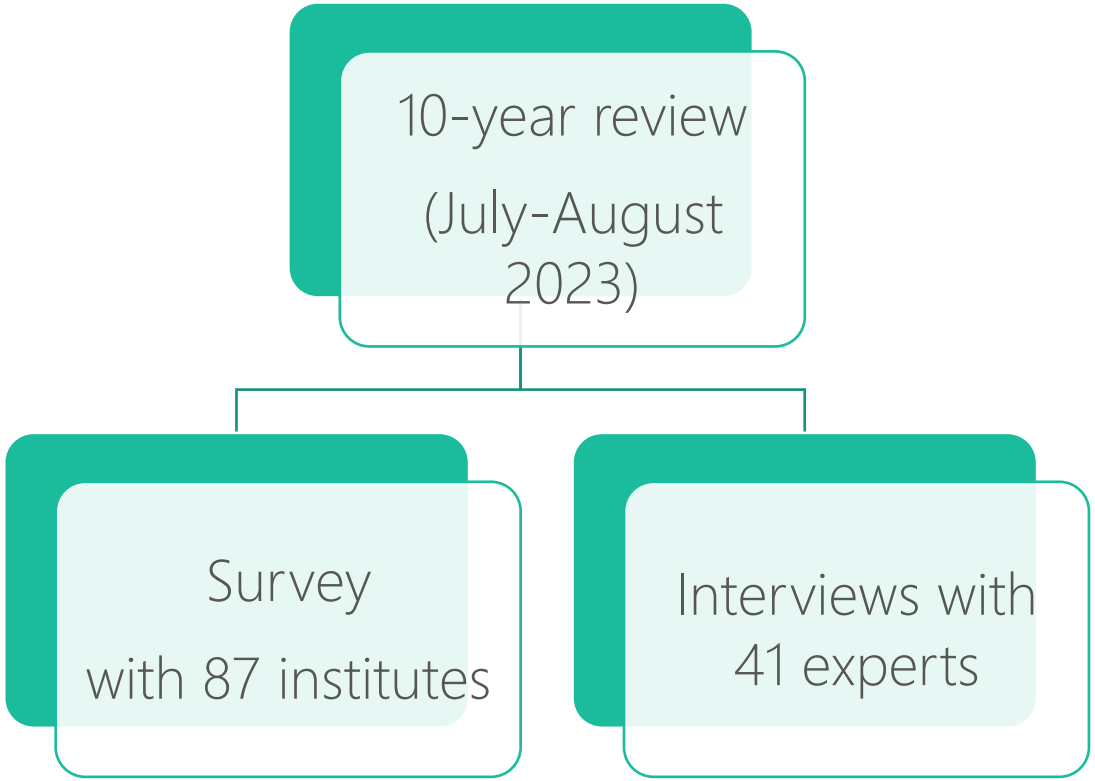


UNBigData

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

10. **High-level goals of multi-year plan**
    Please check all those topics below which are part of your innovation strategy

IT Management

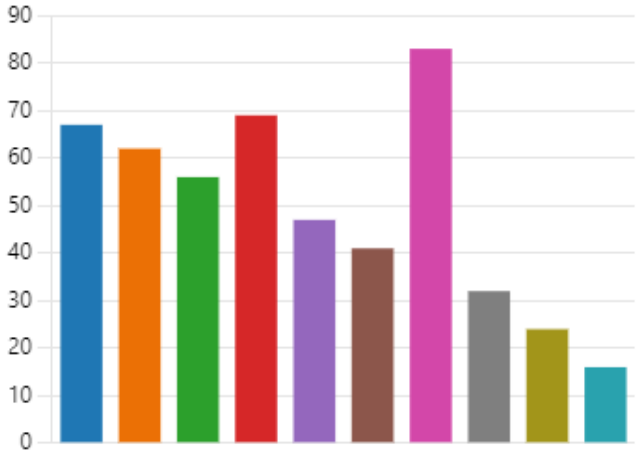
38. **Infrastructure**

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

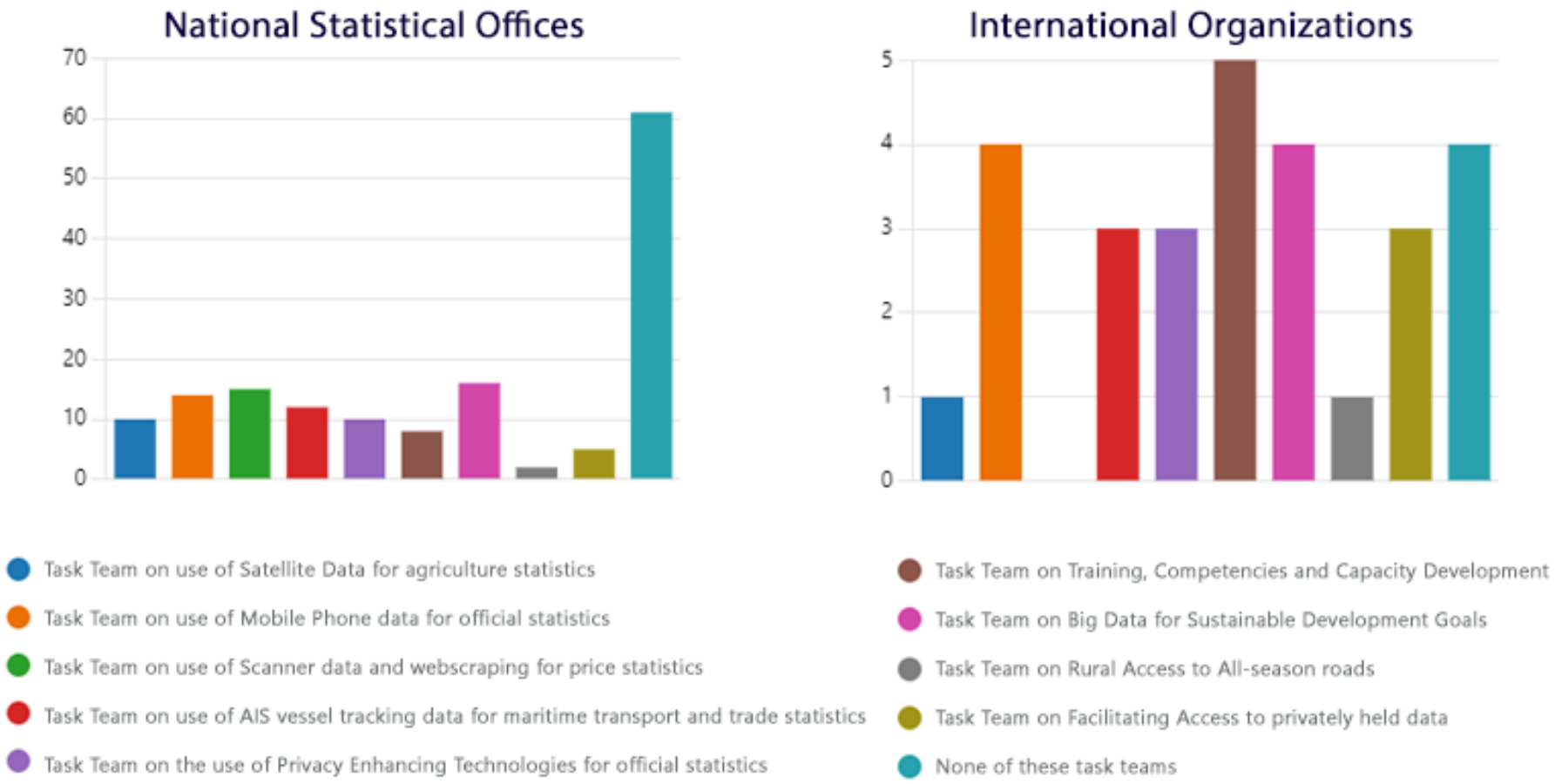
New Data Sources

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

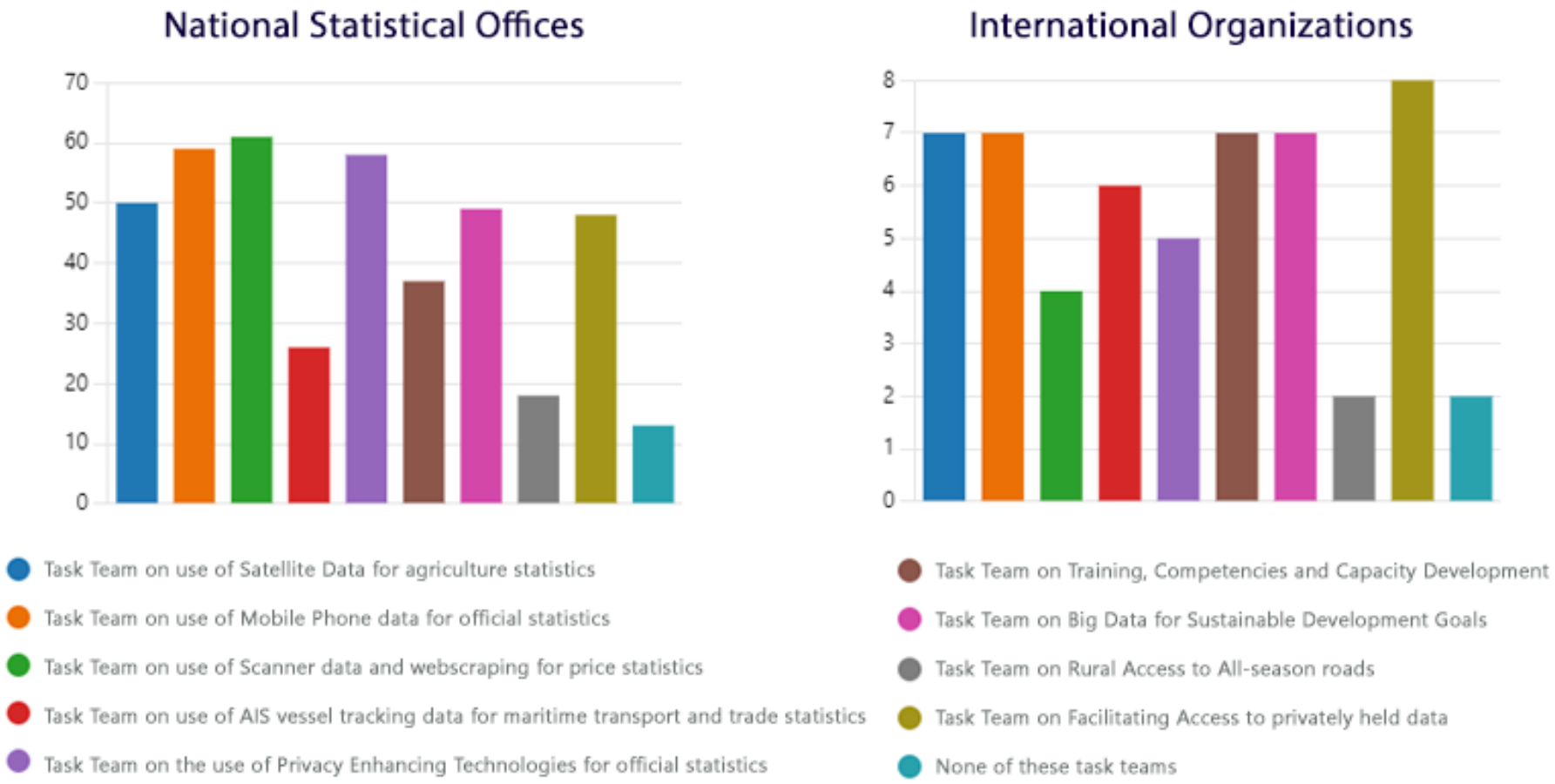


20. **Alternative data sources, including Big Data or webscraping and other privately held data**
What other data sources do you use or are you considering using in the future?

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



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



Summary of Survey Results

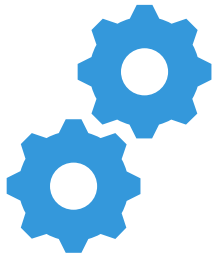
Almost 4 out of 5 NSOs have explicitly incorporated references to modernization, innovation, data science, and the use of big data, into their strategic agendas.

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

Correspondingly, more than 4 out 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