Statistical Commission
Fifty-second session
1–3 and 5 March 2021
Item 3(j) of the provisional agenda
Items for discussion and decision: big Data

Background document
Available in English only

Task Teams of the Global Working Group on Big Data for Official Statistics

Prepared by the Global Working Group on Big Data for Official Statistics

Table of Contents

Introduction

The Global Working Group on Big Data for official statistics delivers most of its work through task teams, which develop methods, prepare handbooks, conduct capacity-building activities, acquire data, make algorithms available in the methods service and demonstrate the active use of the data and services available on the United Nations Global Platform. The Global Working Group has active task teams on the use of satellite imagery data, mobile telephone data, scanner data and automatic identification system vessel tracking data, on big data and the Sustainable Development Goals, on privacy - preserving techniques and on training, competencies and capacity development. A more detailed overview of their work is provided in this background document.

Some new task teams are starting their work, notably a task team on the rural access index and Sustainable Development Goal indicator 9.1.1, which is led by the World Bank, and a task team on the acquisition of global private sector data, which was proposed and approved at the plenary meeting of the Global Working Group in November 2020. This team will approach global companies to negotiate access to their global data sources under global arrangements and will work closely with the co-investment use case on data acquisition, exchange and sharing recommended by the Friends of the Chair group on economic statistics and the proposed United Nations network of economic statisticians.

Members of the Statistical Commission are all invited to participate in any of these task teams.

Task Team on Earth Observation Data for Agriculture Statistics

Part I. General Overview

Members of the Task Team

Chairs: Gordon Reichert and Alice Born (Helene Bérard as of 18 January 2021)

Members:

Alessandra Alfieri (UNSD)	Ronald Jansen (UNSD)
Dominique Habimana (FAO)	Sangita Dubey (FAO)
Pietro Gennari (FAO)	Yakob Mudesir Seid (FAO)
Ivo Havinga (UNSD)	Salar Tayyib (FAO)
Sean Lovell (UNSD)	Sara Burns (Canada)
Jessica Ying Chan (UNSD)	Olga Cowings (Data4Now)
Jillian Campbell (CBD)	Maria Ximena Correa Olarte (Colombia)
Jose Rosero Moncayo (FAO)	Manzul Kumar Hazarika
Lorenzo DeSimone (FAO)	Kerrie Mengersen (QUT)
Andrew Davidson (Canada)	Jacinta Holloway (QUT)
Talip Kilic (World Bank)	A. Aguilar (Mexico)
Tomasz Milewski (Poland)	Benjamin Stewart (USA)
Michael Schmidt (Germany)	Artur Lacynski(Poland)

National Statistical Agencies	Statistics Canada Statistics Poland INEGI (Mexico) DANE (Colombia) Egypt Indonesia Senegal USA (USDA – NASS) Germany (Space Agency)
International Agencies	FAO World Bank UNSD
Other	Queensland University Agriculture and Agri-food Canada Data4Now

A. Objectives of the Task Team

A Task Team on Satellite Imagery was first created in 2014 (under the Global Working Group on Big Data for Official Statistics), with a mandate to identify approaches for collecting representative training data; and develop and implement methods using satellite imagery and the training data for producing official statistics, including the statistical application of predictive models for crop production yields. The task team was later renamed to the Task Team on Earth Observation Data for Agriculture Statistics to not limit the data sources just to satellite imagery.

The Task Team on Earth Observation Data for Agriculture Statistics aims to provide strategic vision, direction and development of a work plan on utilising satellite imagery and geo-spatial data for official statistics and indicators for the UN sustainable development goals. The Task Team is building on precedents to innovatively solve the many challenges facing the use of satellite imagery and geo-spatial data sources. The focus of the Task Team is to:

- Identify reliable and accurate statistical methods for estimating data of interest;
- Suggest approaches for collecting representative training and validation data of sufficient quality;
- Research, develop and implement assessment methods for the proposed models including measures of accuracy and goodness of fit;
- Establish strategies to reuse and adapt algorithms across topics and to build implementations for large volumes of data.

The first focus of the Task Team is to recommend EO data sources, methods and training for producing agriculture statistics, which is a priority for international and national organizations around the world. The second work stream of the Task Team is using EO data for land cover and land use. Both work streams are described below.

a. Agricultural statistics

In 2020, an opportunity was identified for a FAO collaboration with the Task Team on EO Data for Agriculture Statistics. The work includes providing technical hosting support on the UN Global Platform to the FAO, and collaboration on a research project with members of the EO Task Team.

The FAO has leveraged the UN Global Platform using the FAO OCS - European Statistical System (ESS) joint initiative on EO-assisted crop monitoring for the generation of official statistics in developing countries, such as Senegal and Uganda. The project uses optical data from Sentinel 2 and in-situ measurements (crop location, crop type and crop yield) provided by national counterparts under the 50X30 - Agris Umbrella Programme. The work complements the analytical work with the development of best practices/methodology and specific training on EO preprocessing, and the actual analysis (e.g., Random Forest, Spatial Regressions etc.)

b. Land cover and land use statistics

Until now only very few countries in the world produce consistent time series national land cover maps that meet the UN SDG reporting requirements (temporal, spatial and thematic resolution). Furthermore, national products are usually generated using country based standards (e.g. definition of land cover classes and methods to generate these) which compromise international comparability of results, thwarting the full scope of the SDG framework. FAO has developed a standard for land cover classification, the LCCS, which is the de-facto ISO standard, and which has been used to develop the land cover definition under the System of Economic Environmental Accounts (SEEA) of the UN.

The development of a global land cover product, which could support SDG monitoring ensuring standardization and harmonization of national and subnational figures and adequate accuracy (> 90%) remains a void to be filled, and is therefore proposed here as a potential for a collaboration between the FAO and the Task Team on EO. Use of advanced machine learning algorithms, like for example deep learning will be explored given the high potential for such algorithm for accurate land cover classification. Due consideration should be given to the choice of algorithms based on the availability and quality of the training data and on the need to update the land cover in the following years with minimum training data.

Taking advantage of FAO's ongoing collaboration with Space Agencies, a joint collaboration is proposed given that the required resources for such a project would go beyond the technical capacity of the solely UN working groups.

The EO Task Team aims to bring together international and national experts in land cover maps and classifications to advance this project at a future date.

B. Main outputs so far delivered by the task team

In December 2017, the Task Team released a handbook on *Earth Observation for Official Statistics*, which provides a guide for National Statistical Offices considering using satellite imagery data.
Link:
https://unstats.un.org/bigdata/task-teams/earth-observation/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf

The handbook contains a brief introduction on the use of earth observation (EO) data for official statistics, types of sources available, methodologies for producing statistics from this type of data and quality indicators for spatial data. It also summarizes the results of four pilot projects produced by Task Team members: Australian Bureau of Statistics (ABS), Australia, Instituto Nacional de Estadística Geografíca e Informática (INEGI), Mexico, Departamento Administrativo Nacional de Estadistica (DANE), Colombia and Google. Supplementary material includes a methodology literature review, and the full reports of the pilot projects.

The following covers the activities of the Task Team in 2019 and 2020. Over the past year, the Task Team has had the opportunity to re-set its priorities by actively engaging with the FAO and other new members while focussing a few deliverables.

i. Methodology

Leveraging the UN Global Platform – Collaboration between FAO and UN-GP

In April 2020, the FAO and the UNSD have established a joint research and development collaboration on the UN Global Platform using Earth Observation (EO) based tools for producing improved agricultural statistics in countries, and for advancing methodological research on EO applications for crop monitoring and land cover.

The joint workplan for 2020 has focused primarily on the development of an EO cloud computing environment that could support the implementation of the FAO EOSTAT project. The project seeks to build in-country capacity in Senegal and Uganda in the operational use of EO data for the production of official crop statistics employing established methods and tools. Specifically, the Sen2Agri tool box developed by the European Space Agency, in collaboration with FAO, has been identified as the tools box of choice as it allows for the semi-automatic EO data preprocessing and provides a user friendly graphic interface allowing for the actual classification of the EO data into crop type maps.

The Sen2Agri tool box requires specialized IT skills for the deployment, and deep knowledge of the workflow for the proper set up. The actually running of the application requires high storage and high computing power. As a result of the joint collaboration, it was possible to overcome such barriers by deploying the Sent2Agri tool box on the UNGP on one end, and by developing Lambda scripts on the other end to automatize further deployment of the Sent2Agri tool box for other projects.

In April/June 2020, the design of the EOSTAT solution was finalized and a technical document was produced. This was a major milestone. The requirements of the Sen2Agri workflows and the data requirements for the two countries (Senegal and Uganda – e.g. number of Sentinel 2 Tiles to acquire and process) allowed to design a solution that provides enough storage and computing power, optimizes performance/costs, and allows for automatic deployment through Amazon Lambda scripts. The final solution is therefore easy to scale up for more countries.

Below is a comprehensive map story describing the EOSTAT project including Senegal, Uganda, Senegal and recently added, Afghanistan.
https://hqfao.maps.arcgis.com/apps/MapJournal/index.html?appid=f9fd5da0a2d44a40aee4d1cbb83a38ce

ii. Training

The Task Team has created a training sub-task team on the use of earth observation data for agricultural statistics, which is deemed to be a training priority by member countries. The training task team is led by Australia’s Queensland University of Technology and the FAO who provide expertise in training and the use of earth observation data for agricultural statistics. The assignment is backed by many task team members who have provided material and links, fundamental to the creation of the courses. To date, the team has developed a three-stage curriculum which is meant to lead students from introductory remote sensing knowledge to advanced courses with up-to-date data and training sources. The curriculum has been created with the overall goal to teach the fundamentals of using satellite imagery and provide programming skills with relevant use cases for its users.

iii. Events

UN Big Data Conference: Use of Satellite Data for Agriculture, Environmental and Ocean Statistics

On September 1, 2020 the Task Team organized a session on the use of satellite for agriculture, environmental and ocean statistics at the 6th International Conference on Big Data for Official Statistics. The session had seven presentations from experts in the field of remote sensing to show case how this technology has transformed production of statistics for agriculture, environment and ocean science across the globe.

The first presentation from Australia's Queensland University of Technology (QUT) discussed the research on the use of random forest classifiers for classifying missing data from phenomenon such as cloud cover in satellite imagery for land cover mapping. The innovation is meant to be used in changed forest clearing events, and land cover classifications where cloud cover is more frequent.

The second presenter from QUT gave an overview the earth observation analysis work done for official statistics within the EO task team, and its overarching goals to align with the sustainable development goals of 2050.

The third presenter from the USGS provided an overview of ecosystem accounting through the joint efforts with the UN's ARIES (Artificial Intelligence for Environment and Sustainability).

NBS from China outlined the past, current and future projects using planes and drones to capture EO data with a goal to provide country wide land cover and household data sets, in addition to their use of EO data for agricultural statistics.

The FAO demonstrated the use of EO data and development of crop statistics to developing countries, such as the current work in Senegal. The FAO has leveraged the use of the UN Global Platform for development of essential programs and sharing of data for production of crop statistics.

The Vietnam Academy of Forest Science presented a case study of Vietnams' Quang Ninh area oceans. Ocean pollution data, coral reefs maps and hydrology raster's were used with EO data over time to map changes in the ocean environment in this area, which led to environmental protection policies.

Microsoft presented developments in AI focusing on providing education for environmental and ecological changes in response to human impacts, which to date has been collaborative with 90 countries.

iv. Other

New Joint Task Team
At the September 25, 2020 meeting of UN Committee of Experts on Food Security, Agricultural and Rural Statistics (UN-CEAG), the creation of a joint task team on the use of earth observation for agriculture statistics and land cover mapping between UN-CEAG and UN-GWG on Big Data was approved, given the overlap identified between the work programmes of the two task teams.

An official communication between UN-CEAG and GWG-Big Data task team was prepared, and the inclusion of the UN-CEAG members interested in this topic (i.e. Egypt, Indonesia, Mexico, Senegal, USA (USDA-NASS), ADB, FAO, UNSD and the World Bank) will become members of this task team. Under this arrangement, the joint task team will develop and implement joint programme of work that would be monitored by both Experts groups and reported on separately to the UN Statistical Commission. The Committee approved this proposal and elected Canada as the chair of the joint TT since Canada she is already leading the TT of the Global Working Group on Big Data for Official Statistics.

Part II. Ongoing deliverables

i. Projects

A research sub-task team led by Canada is providing expertise in research and remote sensing science to investigate the minimum in-situ data necessary to interpret satellite or other types of earth observation data to produce crop estimation classification for official statistics. The research agenda aims to gather experts who use various methods of remote sensing estimations and classifications.
A research paper will be used to produce a handbook to demonstrate methods to reduce the need for in-situ data, which are often expensive and timely to collect; and provide expertise in data collection and use of EO data for crop statistics. The sub-task team has identified experts to potentially contribute to the following topics: collection of data, supervised and unsupervised classification techniques, and the use of machine learning methods. Further developments of the task team include collection of the references and sources to piece together methods and recommendations. The final product will be openly available the UN Global Platform.

ii. Methodology

The World Bank has recently joined the task team adding to the toolbox of methods for agriculture statistics under the 50 x 2030 initiative, in particular for Malawi. The World Bank and the Task Team are exploring areas of further collaboration.

iii. Training

Further developments of the Training Sub-task Team are planned to include the integration of the material onto the UN Global Platform, in addition to a special project with partner the FAORAP. The project is designed to develop a national project in Afghanistan on crop monitoring with the National Statistics Office. Team will work closely with the Task Team on Training and is planning a side event for the 52nd Session of the UN Statistical Commission (March 2021).

iv. Events

Two side events are planned for the 52nd Session of the UN Statistical Commission:

- Design of an EO Training Programme (11 March 2021);
- Using Sentinel-2 Satellite Imagery for Mapping of Crops in Senegal (23 February 2021)

Task Team on AIS data

Part I. General Overview

– Members of the Task Team

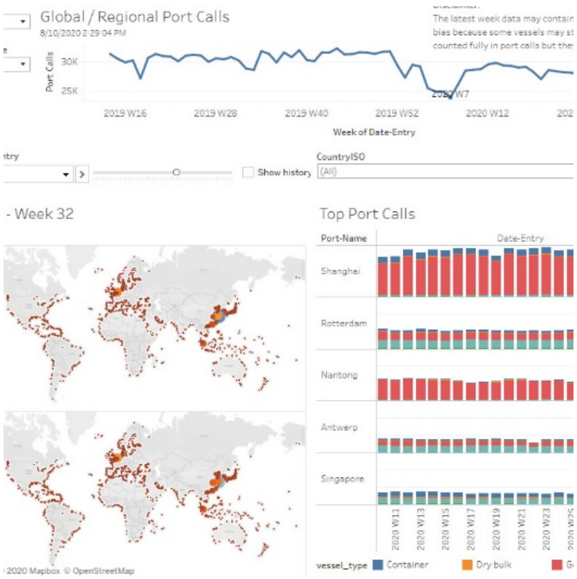
Chair: UNSD

National Statistical Agencies	CBS Netherlands Hellenic Statistical Authority - Greece Maldives Monetary Authority Statistics Denmark Statistics Indonesia CSO Ireland Statistics Netherlands Statistics New Zealand Statistics Poland UK Department For Transport UK Office for National Statistics
International Agencies	ADB Eurostat IMF IMO UN Global Pulse UNCTAD UNSD
Other	Autoridad Marítima de Panamá DFID Data Science Hub Environmental Change Institute, University of Oxford MarineTraffic Oceanbolt The Turing Institute University College London

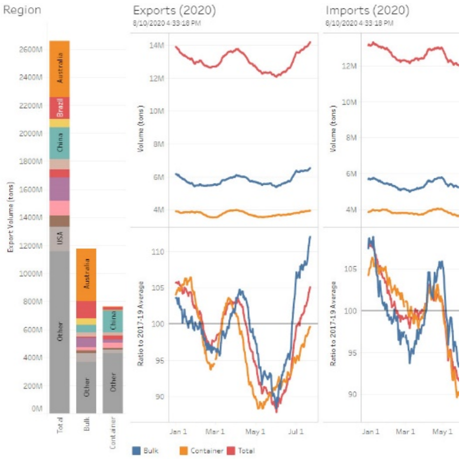
– Objective

The Task Team in AIS data analyses data coming from the automated identification system (AIS), a signaling (sensor) system of vessels to share various information about ships' location, speed, status, and other relevant information. The team aims to develop algorithms and methodologies for measuring freight transportation, traffic within harbors, economic trade indicators, CO2 emission, fishery, and other experimental indicators fit-for-purpose; and conduct training for using AIS data in these domains. The team uses the UN Global Platform for global collaboration on projects involving AIS data (http://location.officialstatistics.org)

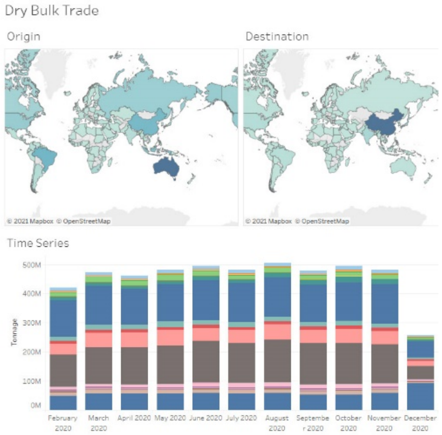
- Main outputs so far delivered by the task team:
  - Methodology
    - Delivery of the Live AIS Handbook (Initially released in February 2020): https://unstats.un.org/wiki/display/AIS/AIS+Handbook+Outline. It consists of the AIS dataset description and an inventory of use cases – including the methodologies applied and the way forward.
      - Faster economic indicators: Time in port and port traffic (UK ONS)
      - Maritime indicators (UNCTAD/MarineTraffic)
      - Official Maritime statistics: Port visits (Eurostat)
      - Completing statistics on Inland waterways (Statistics Netherlands as part of the ESSnet Big Data II: Tracking ships 2018-2020)
      - Mapping fishery activities (EU JRC)
      - Ships in distress (UN Global Pulse)
      - NOx, SOx, and CO2 Calculation (Université du Havre)
      - Nowcasting Trade Flows in Real-Time (IMF)
      - Experimental Statistics of Daily number of a vessel (Statistics Denmark)
      - Real-Time Data on Seaborne Dry Bulk Commodities (Oceanbolt)
  - Projects
    - Release of AIS port calls and trade volume dashboards (August 2020): https://marketplace.officialstatistics.org/ttt-dashboards



The latest week data may contain
bias because some vessels may st
counted fully in port calls but the



- Release of AIS dry bulk commodities trade dashboard (January 2021):



- Training
    - The AIS Data Week (March 2020): https://unstats.un.org/unsd/trade/events/2020/AisDataWeek/default.asp
    -
- Events
    - The AIS Hackathon (September 2020): https://unstats.un.org/bigdata/events/2020/ais-hackathon/
- Other
    - Consistency of the regular task team meetings
    - Increased task team size and number of users at UNGP AIS data platform
    - The highlight of AIS data in UNCTAD Report on Maritime Transport (November 2020): https://unctad.org/topic/transport-and-trade-logistics/review-of-maritime-transport
    - Participation at various events and for a such as ICAO webinar, GWG Big Data Conference, UN World Data Forum "Road to Bern" series in 2020

Part II. Ongoing deliverables

- Methodology
  - Ongoing update on the AIS Handbook by integrating selected outputs from 2020 AIS Hackathon
  - Benchmarking shipments of bulk dry commodity indicators with official customs data in Indonesia
- Projects
  - Operationalizing the output of AIS Hackathon into continuous monitoring activities (i.e., through the dashboard)
  - Cruise Tourism by CBS Netherlands
- Training
  - Providing continuous technical support on how to use AIS through Slack.
- Events
  - N/A
- Other
  - Acquisition of Global Ships Register database in collaboration with IMO
  - To better support simultaneous extraction and improve users' experience, it is planned to upgrade the AIS platform's architecture by utilizing Data Bricks as a front-end analytical platform.
  - Also, due to a change in the funding mechanism, the task team plans to engage with the GWG bureau and relevant stakeholders to ensure sustained funding for the platform.

Part III. Planned deliverables

- Methodology
  - Explore the relevant datasets such as customs, shipping manifest to increase the analytical value of AIS
- Projects
  - Establish a project to monitor the establishment of the green corridor in the Panama Canal to support the IMO's resolution in the reduction of GHG
  - Porting the output of ESS.NET AIS workstreams to the UN Global Platform environment
- Training
  - Releasing self-paced e-learning on AIS foundation course
  - Conducting AIS Data Week in March 2021
- Events
  - Conducting the 2021 AIS Hackathon in September
  - Sharing activities of the task team in the UNCTAD newsletter
  - Regular update of the task team website
  - Contributing to the Big Data Conference, UN World Data Forum in 2021

Part IV. Inventory of Methodologies

Faster economic indicators: Time in port and port traffic (UK ONS)

The ten biggest UK port areas from 2017 were manually defined from a map for each port using the typical berth positions, resulting in rectangular bounding boxes. In the case of the port of Grimsby & Immingham, due to the distance between the sites, two bounding boxes have been defined and the presence of a ship in either of them is considered in-port state. If the ship's location did not fall in any of the defined port bounding boxes then the in-port state was default value 0, defining a group of ships that are out of port. A ship is considered in port if its reported position was inside the port bounding box. The in-port states are marked from 1 to 10, corresponding to the port numbers. The data were grouped by the in-port state of the ship to produce the outputs for each of the ten ports individually and aggregated for all ports. The 'time-in-port' indicator was computed by summing all the periods of the time spent of ships having in-port state corresponding to the port over each port and each month. If a ship's AIS transponder was switched off inside a port, the time in port was counted only if the following message received from the ship was also within the same port. This rule eliminated the outliers in data resulting from moored ships switching off their AIS equipment and later leaving port without reactivating it or from ships that for some reason change their Maritime Mobile Service Identity (MMSI) while in port.

Similarly, the 'total traffic' indicator was computed by grouping the data by port and month for the number of unique ships, based on MMSI number. As the total traffic indicator measured the number of unique ships entering port each month, it is not sensitive to ships that spend very long periods in port, e.g. pilot boats, or to have frequent port calls, e.g. ferries. On the other hand, the 'time-in-port' captures all time that ships spend in port and it may increase relative to the 'total traffic' indicator if either there are delays in port, or it takes longer to upload ships due to more cargo on board.

The data on port activity indicators were compared to monthly economic statistics:

Gross value-added, chained volume measure, seasonally adjusted (source, ONS)
Trade in goods: ONS, imports and exports, current prices defined as change in international ownership in accordance with European System of Accounts 2010 (CP, SA, source: ONS)
UK overseas trade statistics: HMRC, imports and exports defined as the movement of goods across international borders (CP, NSA, source: HM Revenue and Customs)

Maritime indicators (UNCTAD/MarineTraffic)

To calculate the number of port calls, only arrivals are selected. Cases with less than 10 arrivals or 5 distinct vessels on a country level per commercial market as segmented are not included. Passenger ships and RO/RO ships are excluded from the time at port calculations.
Time is calculated as the median time. Due to statistical outliers, the average time vessels spend in port is longer for practically all countries and markets. Ships can spend a long time in a port, for example for repairs skewing the data.

The data comprises 8 markets (based on ship type): passenger ships, wet bulk, container ships, dry breakbulk, dry bulk, RO/RO (roll-on/roll-off), LPG, LNG.
For these markets the following variables are computed:
- Number of arrivals
- Median time in port (days)
- The average age of vessels(each ship is counted as often as it is called in the country's ports)
- Average size (GT) of vessels: (each ship is counted as often as it is called in the country's ports)
- Average cargo carrying capacity (DWT) per vessel(each ship is counted as often as it is called in the country's ports)
- Average container carrying capacity (TEU) per container ship: (each ship is counted as often as it is called in the country's ports)
- Maximum size (GT) of vessels
- Maximum cargo carrying capacity (DWT) of vessels
- Maximum container carrying capacity (TEU) of container ships.

Official Maritime statistics: Port visits (Eurostat)

The first step was to construct a reference frame of ships, for two reasons. Maritime statistics only apply to maritime ships carrying goods, however other ships entering ports also emit AIS signals. Filtering data from other ships reduce the huge amount of data. The other reason is the need for the frame of ships to categorize ships. Maritime ships can be identified on the basis of their so-called IMO number (International Maritime Organization). This number is not included in the dynamic messages (containing information needed to determine the location of a ship), they only contain the MMSI number. These MMSI-numbers also originate from non-maritime ships that should not be included for maritime statistics. Furthermore, not all statistical offices (can) collect MMSI or IMO numbers, some collect call sign as a shipping identifier. Thus, the reference frame of maritime ships should include all three ship identifiers: MMSI, IMO and call sign. Still, AIS data does not provide sufficient information for a complete reference frame of maritime ships as maritime statistics require more specific information on the identity of vessels. For example, the type of vessel defined in AIS data is less detailed than required by maritime statistics based on Directive 2009/42/EC. Therefore, in order to get more detailed ship information, AIS data has to be linked to existing ship dictionaries which was not present.
The reference frame was constructed by linking all MMSI-IMO couples in the static messages. First, couples with invalid IMO or MMSI numbers were filtered om the basis on length for MMSI and for IMO the check digit.

To generate port visits on the basis of AIS data, the reference frame was used to investigate which ships had been in the port on one day for Poland (Świnoujście) and the Netherlands (Amsterdam). For MMSI's in the reference frame, it was checked whether their location (latitude and longitude) were present in a certain area. Statistics Poland did this for one day for the external boundaries for the port of Świnoujście Statistics Netherlands did this for the port of Amsterdam using a simple bounding box. Ships that were identified were compared to ships from the port statistics.

Completing statistics on Inland waterways (Statistics Netherlands as part of the ESSnet Big Data II: Tracking ships 2018-2020)

In the Dutch statistical process of IWW, the operationalization of a journey is a movement between two locations where loading or unloading takes place, or the location where a ship moves outside Dutch inland waterways. To complete information on missing journeys, information from AIS can be used. To link AIS data to the IWW statistical data, the same concept of a journey is applied for AIS data. The method comprises 3 steps: preprocessing, deriving journeys and linking AIS and traditional journey information.

Preprocessing

From the AIS data, dynamic and static messages are separated. For each ship, a file is constructed with all dynamic messages and one file with all static messages. The dynamic messages contain information such as the ship's location and speed. The static messages contain information on the ship characteristics such as identity, size and ship type. The dynamic information is filtered to contain messages where the ship's speed is faster than 0.2 knots and latitudes and longitudes are in the Dutch range. Navigational status was not used, because these were not always filled in or not correctly filled in. In addition, messages are filtered that do not have valid MMSI (i.e. not having 9 digits). The file with static messages is deduplicated on the basis of ship type (as a ship might have different uses through time). Then, the single static message is added to the dynamic messages.

Deriving journeys

The operationalization of a journey is the movement of a ship between two locations where goods are (un)loaded or the point where a ship moves outside Dutch IWW. For this, each file containing information per ship is used. When the time interval between two successive data points is over an hour or the difference in location is larger than 200 meters, the first data point is considered to be the end of a journey and the succeeding data point the start of a new journey. Each location is linked to a register of terminal locations using the nearest node method. As the definition requires locations where ships (un)load information, the location register also has to contain locations where ships regularly stop for other activities such as waiting areas for locks or sleeping places. This enables differentiating locations for (un)loading conditions and other stopping reasons. This process requires a good register with terminals

Linking AIS and traditional journey information

Journeys from AIS (lasting from tA1-tA2) were linked to "lock" journeys (lasting from tB1-tB2). The following criteria were used: the starting time of the "lock" journey had to be before the end time of the AIS journey and end time of "lock" journey had to be after beginning time of AIS journey:tA2 > tB1 and tB2 >t A1. Then, for the "lock" journeys the amount of overlap with the AIS journey is calculated using the following link index:

$$\text{Link index} = ((tA2-tB1))/((tA2-tA1))*((tB2-tA1))/((tB2-tB1))*(tB2-tB1)$$

If the time matches, link index will be around a value of 1, otherwise the value will lower. Values with the highest link indices are matched. AIS journeys that are not linked to “lock” journeys are taken to be missing journeys from the lock data.

Mapping fishery activities (EU JRC)

First, the data was cleaned. For this, the position and speed data cleaned and filtered to an interval of 5 min between consecutive observations (reducing 120 mln to 60 mln messages). Messages with zero-velocity messages relating to periods when the vessel is likely to be in a port were excluded. Resulting speed profiles were analyzed. This resulted in speed profiles with a bi-modal distribution, where corresponding to fishing behavior is extracted. This algorithm is based on the assumption that speed during fishing activities is lower than during steaming (Mazzarella et al., 2014).

Note that, factors such as vessel size, area, and fishing gear result in specific mean and standard deviation values of the speed bi-modal distributions. For this reason, the identification of fishing behavior has to be implemented for each individual vessel (Natale et al., 2015). For the map, aggregate results were turned into density maps: the resulting points classified as fishing were aggregated into 1 km2 cell.

Ships in distress (UN Global Pulse)

Information from the different sources on rescue sequences is combined into a so-called quantified rescue. This approach is used to unify narrative threads from a variety of sources, tying the observable physical traces of a rescue operation to qualitative sources of information on what happened and how events unfolded. This reduces “subjectivity” in the data. To produce a quantified rescue, incident descriptions were used to identify the ships involved in a rescue. Then, timestamps were used to link these descriptive statements or tweets to the AIS coordinates that are closest in time. This is not always possible as descriptive data can be vague or incomplete. Another problem was that for some AIS data the frequency was too low. As rescues involve sudden movements, the interval between two data points cannot be too long, otherwise, deviations in trajectories cannot be picked up. Also, the amount of available information varied substantially by the incident.

To produce a large-scale training dataset for machine learning, 77,372 points were manually geotagged, spanning four search and rescue ships over a period of 100 days. Points were labeled according to two estimated activity types—“Rescue” and “Non-rescue”—based on the shape of the trajectory and the corresponding rescue ship’s speed. Models of rescue behavior were trained to predict this binary outcome based on the characteristics of trajectory points, including speed, course over ground, latitude, longitude, day of week, hour of day, and month of year. Two approaches were taken:

First, classification algorithms were trained on the raw point dataset.
Second, classification algorithms were trained on clusters of points. Specifically, points were clustered prior to classification using the Cluster-Based Stops and Moves of Trajectories (CB-SMoT) algorithm (Palma et al. 2008). Density is defined using two parameters: eps, a distance parameter, and mintime, a time parameter. CB-SMoT uses these parameters to classify points into three categories: (1) Core points, from which the object travels less than the distance threshold eps in either direction within a period of

length mintime; (2) Border points, falling within eps of a core point; and (3) Other points, which are neither border nor core. Together, core and border points form clusters that represent periods of slow motion. Although the algorithm is designed to find stop points, it was used here to break trajectories into segments with different motion profiles by allowing sequences of “other” points to form their own clusters.

Automatic characterization was applied using standard binary classification algorithms—AdaBoost, Support Vector Machines (SVM), and logistic regression—with 10-fold cross-validation.

NOx, SOx, and CO2 Calculation (Université du Havre)

The authors identified three structural variables and one geographical variable which can be used to calculate emissions from maritime traffic. They are as follows:

Gross tonnage (DWT): implying the need of energy for moving the vessels

Age of vessel (Ag): recently constructed vessels have a more efficient propulsion system or equipped by “Dual Fuel”. It is categorized into three periods: before 1990, between 1990 and 2005 and after 2005)

Speed of vessel (SOG): which has an impact on fuel economy (i.e., “Slow Steaming” could reduce fuel consumption up to 50%)

Geographical concentration of traffic

The variables are weighted based on their impact on the production of emissions (Ag=3, DWT=2, SOG=1). The formula of emission potential is as follows:

((Agx3)x(DWTx2)xSOG)/Ut

With Ut representing a unit of time in observation

Nowcasting Trade Flows in Real-Time (IMF)

The methodology used on the study follows two steps:

Cargo ships are identified by a filter and static and voyage-related information for the identified ships is aggregated:

The filter identifying the cargo ships follows three rules: 1) Bunkering tankers providing fuels to vessels located at seaports, 2) ships arriving but not departing, and 3) ships that stay in the port boundaries only for a short time or for too long are omitted.

High-frequency indicators are derived: On a weekly basis, the cargo number indicator that counts the number of incoming ships and the cargo load indicator based on information on the ship's deadweight tonnage and the reported draught are calculated.

To verify the indicators, they were compared with the official Maltese statistics.

Experimental Statistics of Daily number of a vessel (Statistics Denmark)

Statistics Denmark has published index of daily number of vessels visiting number Danish ports using Danish AIS-data. It is accessible from https://www.statistikbanken.dk/aisdag. The data processing steps consists of the

following
1. Data reduction
2. Selection of arrival and departure observations
3. Delimitation of ports
4. Linking arrival and departure observations with ports and creating statistics.

Data reduction

The first step is data reduction. The sole purpose is to reduce the amount of data that has to be processed, and it consists of three parts: geographical reduction, reduction of observation frequency and reduction of analysed vessel types.

AIS data contains all the information from AIS that the Danish receivers have registered. That includes multiple observations that do not relate directly to Denmark but rather Germany, Norway or Sweden. The method used for geographical delimitation is simple as data is delimited to be within a square that covers all of Denmark and, consequently, the southern part of Sweden also. The Swedish ports will be excluded later on in the process. Here it is possible to reduce the observations to ones that are within the Danish waters.

The ships transmit signals at an interval, which is determined by speed and type of activity as well as the type of transmitter/transponder. The most frequent signals are received at intervals of a few seconds. However, since the objective is to examine port calls that level of data frequency is not necessary. Consequently, data is reduced to the first observation per minute.

The activity in the ports can cover many types of activities. In order to support existing port statistics that are centred on cargo handling in the ports, data is reduced to vessel types that are only used for cargo transportation: freight and container vessels.

Determination of arrival and departure observations

The next step in the process is to identify the ports calls with arrivals and departures. Each ship leaves behind a number of position data, and the objective is to identify the two observations, one that represents the arrival at a port and another that represents the departure from the port.

Fundamentally, the process is simple: The data set is sorted by identification of ship and time. Subsequently, all observations where the ship shifts from movement to a standstill are marked (where the navigational status shifts from “under way” to “moored” and the ship goes from moving (more than 1 knot) to (almost) a standstill). These are potential arrivals. The same approach is used for potential departures where the ship is moving and the navigational status shifts from “moored” to “under way”. All of the potential arrivals and departures are matched. For the majority it correlates well, but there are arrivals without departures and vice versa. Possible explanations are:

- Lengthy stay in a port: The average stay in a port is approximately 12 hours, but if the stay lasts for a number of days it can result in one of the matching observations missing, in the beginning or the end of the observed period. This is partly made up for by using data that falls outside the period, e.g. by not preparing the statistics for last month until 5-6 days after the cut-off-date. Long-terms stays in a port is rarely connected with transportation of cargo, but rather a need for repair work or something else, and they are probably not important in relation to the objective of the statistics.

- Turned off transponder: It is not (normally) illegal to turn the transponder off, and a possible scenario is that the transponder is not turned on until the ship reaches the port or that is turned off after arrival at the port, and the crew forgets to turn it on again after departure.
- Data disruption: Data disruption can occur when data from a single costal AIS receiver is not registered, data is not streamed from the Danish Maritime Authority or in cases where data is not collected and stored by Statistics Denmark. The latter was mainly an issue in the initial phase and disruption has been reduced significantly over time.

Delimination of ports

The third step is to exclude the observations that are not actually port calls. At this point, there is still a large number of arrivals/departures not close to a port. Furthermore, data should only include Danish ports. Observations from others ports, especially Swedish, are still a part of the data basis, including the largest cargo port in the Nordic countries, Göteborg. If data is reduced to Danish waters to begin with, this will no longer be an issue.

Once this third step has been carried out with, no less than, a couple years of data (statistics for a single month can be prepared without reducing the port data again. This reduces the monthly production time significantly and new data can be included in the reduced data after the production of the statistics.

The process of this step consists of three parts:

- Gather observations (arrivals are used) in groups (clusters) based on the individual distance between the observations and the number of observations in close proximity.
- Make polygons that cover all the observations in the same cluster.
- Connect the individual polygons to actual ports.

The result of these three steps is a spatial look-up table that consists of polygons and the matching port. A port may be connected to a number of polygons whereas a polygon can only be connected to one port. If an arrival is located within a given polygon, we can then tell which port the port call belongs to. For larger ports, each polygon will typically represent a particular quai.

The first step is to calculate all the individual distances between the arrival points, and subsequently, the arrivals are gathered in groups. If the distance between two observations (irrespective of time) is less than e.g. 50 metres, they are connected. If a third observation is less than 50 metres from one of the first observations, it also becomes part of the cluster. All groups with more than e.g. 5 observations then become final groups or clusters. The number of observations and distance between them are parameters that can be adjusted. The less observations used, the larger the distance should be and the smaller the number of observations in a group should be. With three years of data, 70 metres and 5 observations are used. After the first step, all of the observations have been connected to a group (cluster). All the observations that do not meet the criteria are given cluster number -1 and are considered as invalid port calls.

In the second step, polygons that encircle the individual clusters are made, so that all observations in the same cluster are within or on the edge of the polygon. By looking at the location of these clusters, you will, without further processing, get a good idea of where ports are located. In countries where unofficial ports are established, these ports can also be identified.

In the third step, the individual polygons are connected to a port. This is done in an iterative process in which the basis is a centroid for the port (or something similar – most often, the port coordinate that is on the UnLocode list is used, but it can also be found by a simple search on e.g. Google Maps). The iterative process ensures that no clusters are connected to more than one port and that the polygon is connected to the port that is closest to the polygon. Separation of ports that are close to each other can result in a little extra manual processing, i.e. adjusting the reference points so the individual clusters are connected to the right port. Iteration is done by making a circle that gradually increases around the port's reference point. If a circle overlaps a cluster polygon, this cluster is connected to the port in question, and it cannot, subsequently, be connected to other ports. The maximal distance from the port's reference point to the clusters has to be defined by actual data. In the Danish data, the limit is based on a defined number of clusters for ships that are lying in a roadstead off the coast of Skagen. They are not considered to be lying in port and the limit is set so that these clusters are not included. Since the basis is Danish ports, other ports, primarily Swedish ones, are excluded in this process.

Finally, we end up with a look-up table, in which further information about the individual ports can be found, e.g. municipality, Unlocode, name, business identification number and coastal zone.

Linking arrival and departures with ports

The last step in the entire process consists in matching the port calls (referred to as arrivals and matching departures in step 2) with the polygons that define (part of) the ports. The result is a data set that contains information about the port call (primarily time, identification of the ship and the ship's name) and the port in question (primarily the port's name, code and region). Data can be supplemented with more detailed information about the ships from ship registers, e.g. size, flag, more precise vessel type and owner/operator.

From this point on, the process for production of statistics is the same as usual. Possible sub-classification, indexing and seasonal adjustment and tabulation.

Real-Time Data on Seaborne Dry Bulk Commodities

Raw AIS data is collected from third-party data providers and is processed through Oceanbolt's geospatial algorithms. Our geospatial algorithms model vessel behavior and voyage history. A raw AIS data stream contains two types of information: 1) dynamic data about a vessel's location, speed and direction, and 2) static voyage-related information such as destination, eta, and draught. We make use of both static and dynamic AIS data in our algorithms.

The geospatial processing happens through matching the AIS signals with our polygon database using spatial joins to establish when a vessel enters or exits a polygon (such as a port or a berth). From these events, we are able to generate synthetic voyages. The polygons in our database are annotated with a large amount of metadata such as commodity, terminal operator, trade flow direction (export/import). We use the polygon metadata and our vessel database in combination with the AIS-generated entry/exit events to capture commodity and volume information. Using polygons at berth-level granularity we are able to capture voyage and cargo information all the way down to the individual berth/terminal and this information forms the basis for our cargo prediction models.

To verify the results of the model, we have verified the output data generated by our model with the official trade statistics provided by UNCTAD in addition to various national customs agencies.

Task Team on Privacy Preserving Techniques

Part I. General Overview

– Members of the Task Team

Chair: Matjaž Jug (Statistics Netherlands)

National Statistical Agencies	Australia, Canada, The Netherlands, United Kingdom, United States
International Agencies	Agency names: UN Statistical Division
Other	Names of institutes or companies: Boston University Galois Georgetown University Harvard University Hikari-law Microsoft National Institution for Transforming India Oblivious.ai Open Corporates OpenMined Sarus.tech University of Oxford

– Objective

The Privacy Preserving Techniques Task Team (PPTTT) focuses on 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. Our focus is on methods that enable protecting privacy of data while it is being processed rather than while it is at rest on a system or in transit between systems. The work is intended for use by statisticians and data scientists, data curators and architects, IT specialists, and security and information assurance specialists, so we explicitly avoid cryptographic technical details of the technologies we describe.

– Main outputs so far delivered by the task team:

The task team has been active since April 2018. So far it has released the UN Privacy Preserving Techniques Handbook introducing and describing various privacy preserving techniques, and has started on a second handbook, which will link more to use cases in the statistical community. It is also looking into the legal aspects of using these techniques as well as learning materials for Training programme to increase awareness and help develop skills in statistical community.

- Methodology
  March 2019: Delivery of the first UN Privacy Preserving Techniques Handbook covering five emerging techniques that support protection and sharing of sensitive information:
  - Secure Multiparty Computation,
  - (Fully) Homomorphic Encryption,
  - Trusted Execution Environments,
  - Differential Privacy and
  - Zero Knowledge Proofs

  Handbook can be found at: https://marketplace.officialstatistics.org/privacy-preserving-techniques-handbook
  February 2020: establishment of the Legal subgroup: members have substantial experience and expertise in legal or in encryption techniques, algorithms and relevant products/services. The subgroup is working on the report of legal implications of Privacy Preserving Techniques and development of legal guidelines for organizations implementing these techniques in practice.
  June 2020: Refresh and expansion of membership needed to support new goals and new subgroups to develop second Handbook that will focus on use of Privacy Preserving Techniques in statistical use cases.

- Projects
  April 2020: establishment of the PPT Covid-19 subgroup: this subgroup has been used as a platform to exchange knowledge and experience for urgent development of Covid-19 contact tracking applications and other Covid-19 related use cases during the early stage of Covid-19 crisis. It provided link between projects and expert community. Learning from these use cases will be documented in the future Handbook.

- Training
  September 2020: establishment of the PPT Training subgroup with focus on the development of learning materials, followed by the partnership agreement for collaboration with Openmined.org to create, launch, and operate a series of courses.

- Events
  Presentation of task team activities at workshops and conferences (for example UN Big Data Conference)

Part II. Ongoing deliverables

– Methodology

Legal subgroup is developing Legal Guidelines for Privacy Preserving Techniques with three concrete objectives:

- demonstrate that properly-designed policy objectives regarding data analysis with “privacy by design” throughout the entire data lifecycle can support and incentivize socially-desirable outcomes and organizational decision-making.
- explore and document the extent to which technical guarantees provided by privacy preserving techniques (PPT) can contribute toward compliance with statutory and regulatory privacy requirements.
- explain how PPTs can assist governments and businesses in providing new affordances with sensitive data while adhering to legal obligations and risk assessments.

The document will be accessible and informative to a multi-national audience of policymakers, lawyers, statisticians, data scientists, data curators, IT specialists, and security and information assurance specialists from both the public and privacy sectors.

Second key deliverable of the Task Team in 2021 is Privacy Preserving Techniques Handbook V2. This document will be major refresh of the first Handbook published in 2019 and will have the following focus areas:

- Problem Definition: new Handbook will further explore privacy goals for Statistical Analysis and motivation for the use of Privacy-Preserving technologies. This will be supported by practical use cases.
- Methodologies & Approaches: five privacy-preserving techniques described in the first Handbook will be updated with additional techniques for example Federated learning approach applicable for use of modern AI/ML methods.
- Case Studies in Official Statistics: task team will investigate and document key findings from selected proof of the concepts, pilots or implementation projects in national and international statistical organizations.
- Standards: privacy-preserving techniques is fast evolving field so new Handbook will provide extensive list of relevant standards including some standards that are still in development.
- Legal / Legislation will emphasize main legal/regulatory considerations that must be taken into account when implementing privacy-preserving technologies. Detailed analysis and guidelines will be in the separate document developed by the Legal subgroup.

- Projects

The work of Privacy-preserving Techniques Task Team is closely coordinated with the activities of UNECE Input Privacy Preserving project. This project is collecting and investigating relevant use cases in participating statistical organizations.

- Training

Openmined.org announced 4 initial courses (called "The Private AI Series") to teach privacy enhancing technology:

  - Our Privacy Opportunity (released): a layman-focused course introducing structured transparency, input privacy, output privacy, input verification, output verification, flow governance, and using them to map privacy enhancing technologies onto a series of real-world use cases.
  - Foundations of Private Computation (expected release: February): a technical course introducing the intuitions, maths, and tools of major privacy enhancing techniques: Federated Learning, Split Learning, Intro to Cryptograph, Public Key Infrastructure, Homomorphic Encryption, Secure MPC, Private Set Intersection, Secure Enclaves, Zero-knowledge proofs, and Differential Privacy.
  - Federated Learning Across Enterprises (expected release: Summer 2021): a course on how to deploy federated learning within a real world enterprise environment.
  - Federated Learning on Mobile (expected release: Summer 2021): a course on how to create smartphone apps which learn using federated learning.

More on the courses can be found at: courses.openmined.org

- Events

Task Team will participate in “Preserving Privacy” session of the Conference on New Techniques & Technologies for Official Statistics (NTTS 2021) with two presentations.

Part III. Planned deliverables

- Methodology

In 2021 we will publish second UN Privacy Preserving Techniques Handbook that will focus on applicability of privacy-preserving techniques to concrete statistical use cases. The findings in this Handbook will be deducted from the case studies that are taken from active research at national statistical institutes and international organizations. Furthermore, we will investigate new techniques; evaluate, monitor and report on relevant standards and frameworks and develop guidelines on the legal aspects of privacy preserving techniques.

- Projects

  In 2021 we plan to work with relevant projects for example the UNECE Input Privacy-preserving techniques project to collect information about practical use cases and learn about implementation challenges and solutions. Findings will be documented in the new Handbook. Task team members are also available to provide advice on privacy-related matters such as and privacy-preserving techniques to UN Global Platform projects and task teams that work with privacy sensitive data.

- Training

  Task Team will continue supporting the development of training, exams and certifications for individuals in privacy techniques with the objective of candidates being able to develop methods and procedures for securely processing and exchanging proprietary and sensitive information.

- Events

  In 2021, we plan to organize session specifically dedicated to privacy-preserving techniques at the 63rd ISI World Statistics Congress 2021 scheduled for 11-16 July 2021 (pending final approval by ISI programme board in March 2021). We also plan to prepare e-learning materials and work together with academia, open-source community and other partners to increase awareness and develop skills needed for effective use of privacy-preserving techniques in statistical community.

- Other

  Task Team will facilitate sharing of information and maintain international community about privacy-preserving techniques. In 2021, we plan to regularly share activities in UNGP newsletters, slack channel and regularly update Wiki site and the website of the Task Team. The contribution to Big Data Conference, UN World Data Forum and other relevant events will be also part of outreach activities.

  https://unstats.un.org/bigdata-new/task-teams/privacy/index.cshtml
  https://unstats.un.org/wiki/display/UGTTOPPT/UN+GWG+Task+Team+on+Privacy+Preserving+Techniques

Task Team on Training, Competencies and Capacity Development

Part I. General Overview

– Members of the Task Team

Chair: Statistics Poland

National Statistical Agencies	Canada Denmark Indonesia Philippines Poland Rwanda Switzerland United Arab Emirates United Kingdom
International Agencies	AfDB CARICOM Eurostat GCC-STAT GPSDD ITU PARIS21 SIAP UNSD UNESCWA UNECE UNESCAP
Other	Statistical Society of Canada Canada School of Public Service

The Task Team on Training, Competencies and Capacity Development works on understanding and proposing solutions to help build capacity for institutions that are embracing, or consider embracing, the use of big data in official statistics. The Task Team undertakes projects to understand where national statistical offices currently are on their individual big data journeys, as well as initiatives to understand current and future development needs related to the implementation of big data in official statistics. The work aims to ensure NSOs around the world are increasingly equipped to work effectively with non-traditional data and to produce statistics that are trustworthy, provide value and insight. In addition, the Task Team's goal is to support other UN GWG task teams in the creation of their training programmes by providing guidance on common approaches to the development of training courses.

Main outputs so far delivered by the Task Team on Training, Competencies and Capacity Development:

- Projects:

  - Conduct of a *Global Assessment of Institutional Readiness for the Use of Big Data in Official Statistics*. It presents findings from an assessment of 109 statistical organisations' readiness, considering factors such as strategic data science coordination, legal framework, IT infrastructure and human resources. The final report was presented to the UN Statistical Commission at its 51st session in 2020..

    The results of the assessment are available here:
    https://unstats.un.org/bigdata/task-teams/training/UN_BigData_report_v5.0.html

  - Development of the *Competency Framework for Big Data Acquisition and Processing*. The framework is aimed at NSOs for recognising the expectations and demands for the use of big data in official statistics. It covers the array of skills and knowledge considered relevant for those working with big data. The framework provides guidance to NSOs for assessing big data knowledge and skills gaps, undertaking recruitment for big data posts, and developing staff capability for working with big data.

    The Competency Framework is available here: https://unstats.un.org/bigdata/task-teams/training/UNGWG_Competency_Framework.pdf

  - Development of the *UN Big Data Maturity Matrix* which enables NSOs to undertake a self-assessment of their current level of big data maturity. The matrix aims to help NSOs consider, prioritise and plan for their needs, with the goal of increased capability and confidence in working with big data, and the reassurance and subsequent trust that this cultivates.

    The Maturity Matrix (in PDF format) is available here:
    https://unstats.un.org/bigdata/task-teams/training/Big Data Maturity Matrix v1.0.pdf
    An online version of this matrix is in preparation.

- Events

  - Organization of a session at the 6th International Conference on Big Data for Official Statistics, addressing different aspects of planning and executing training activities for Big Data. The session was the highest ranked of the Conference.

Part II. Ongoing deliverables

- Projects

- Development of an overarching *Big Data Training Curriculum*. The curriculum will map to the *Big Data Competency Framework* and establish global training requirements. It will be based on modern standard taxonomies for defining training levels, etc. Included here is the development of model curricula for Big Data-related training courses to support the work of other task teams in their training activities.

- Development of a *Big Data Training Catalogue*. The catalogue will outline the training courses that already exist, as well as other materials and tools which might help an NSO to improve their big data capabilities. This will be developed in close cooperation with other UN GWG task teams. The catalogue will be reviewed as it grows over time and will later include evaluation statements and classifications linking it to the maturity matrix criteria.

- Completion of *the online version of the UN Big Data Maturity Matrix V.1.0* and planning for an extended version 2.0. The matrix will be made available to countries as an online application for self-assessments. In these self-assessments, NSOs can identify their stage of big data maturity in relation to detailed components/ dimensions of the use of big data, such as legal framework, IT infrastructure, human resources and big data applications in the production of statistics, generating an overall picture, comparing their current stage to their set goals. The envisioned version 2.0 will provide a higher granularity of questions/evaluation criteria and improved linking to existing training courses in the catalogue, as well as supporting guidance materials.

- Managing a Learning Management System (LMS). The LMS will be made available to all task teams to host their online training programmes. The LMS will use the UN Global Platform to provide access to all countries. The role of the UN Global Platform will also be considered for the storage of training products, with links to NSO case studies and big data applications. Courses will produce UN/GWG certificates, if needed.

- *Guidance for developing online training courses*. The Task Team will develop a set of guidance materials (primarily for other task teams) for the development of training materials and online training courses, addressing needs assessments based on existing tools, and requirements for course development at different levels (awareness, beginner, practitioner). This will provide guidance, support and training for other task teams (where required) on how to set learning objectives at the different curriculum levels and sharing good practice in course design. The materials will also include guidance for developing courses on the common platform (LMS).

– Events

- Co-organization of the side event at the 52nd session of the Statistical Commission

Part III. Planned deliverables

- The following projects are expected to start in the second half of 2021:

  - Evaluation of training materials' adherence to guidance. This will ensure that training materials developed in the GWG meet the appropriate level and learning objectives and follow recommended practices of course design.

  - Provision of course evaluation surveys and templates. The development of these materials is in support of tasks concerning preparation of a *Big Data Training Catalogue* and *Guidance for developing online training courses*, allowing the categorization of existing training courses in a consistent way, as well as providing tools for feedback on training courses to developers.

Task Team on Mobile Phone Data

Part I. General Overview

Use of mobile phones, including smart phones, in both developed and developing countries, and in urban and rural areas has been increasing in the past decades. There is a need for detailed statistics for policy and other purposes, including data to monitor the goals and targets of the 2030 Agenda for sustainable development. The increasing data needs also give a challenge to the national statistical offices and generates demand to seek for new data sources, like using mobile phone data. The Task Team on Mobile Phone Data was established to answer for this call in 2015. The Task Team is composed of more than 50 members representing national statistical offices, line ministries, national regulators, academia and the private sector.

Using mobile phone data the task team is currently developing methodologies that can be used to get complementary information and provide quality checks for the indicators included in the SDG monitoring framework, including indicators on the monitoring of orderly, safe, regular and responsible migration and mobility of people (Target 10.7), indicators related to tourism as an enabler of economic growth and creator of jobs (Target 8.9), as well as indicators that will show access and use of ICTs by individuals (Target 17.8 and Target 9.c). Further, based on the methodologies, the task team will develop capacity development courses, including e-learning courses, and will conduct regional workshops and trainings, awareness raising events and projects that could illustrate how countries could benefit from the use of mobile phone data.

– Members of the Task Team

Chair: Esperanza Magpantay, Senior Statistician, International Telecommunication Union

National Statistical Agencies	Brazil, Colombia, Georgia, Indonesia, Italy, Japan, Korea, Mexico, Netherlands, Oman, Philippines, Saudi Arabia, United Arab Emirates,
International Agencies	Eurostat, EU JRC, ITU, IMF, IOM, UNFPA, UN Global Pulse Jakarta Lab, UNSD, World Bank,
Other	Flowminder, GSMA, Positium, Telenor, University of Tokyo

– Objective

Given the wide-spread use of mobile phones in both developed and developing countries, and in urban and rural areas, and against the background of a need for detailed statistics for policy purposes, including data to monitor the goals and targets of the 2030 Agenda for sustainable development, this task team aims to develop methodologies to get complementary information and develop data quality checks for SDG indicators using mobile phone data. These data could be used in areas such as the measurement of the information society, monitoring of orderly, safe, regular and responsible migration and mobility of people, as well as to monitor tourism as an enabler of economic growth and creator of jobs. Further, based on the methodologies developed, the task team will conduct capacity building workshops and trainings, awareness raising events and projects that could illustrate how countries could benefit from the use of mobile phone data.

- Main outputs so far delivered by the task team:
  - The Task Team delivered in 2017 a Handbook on the use of mobile phone data for official statistics describing applications, data sources and methods, and deals with access to mobile phone data through partnership models with mobile network operators. The handbook is available at https://unstats.un.org/bigdata/taskteams/mobilephone/MPD%20Handbook%2020191004.pdf.
  - Task Team conducted workshops in Colombia (2017), Rwanda (2019) and Indonesia (2019).
  - Task Team held an International meeting on measuring human mobility in Georgia (2019).
  - In response to the COVID-19 pandemic, a paper was drafted outlining the guiding principles on the use of mobile phone data for mobility indicators for COVID-19 policy purposes.
  - Members of the Task Team organized sessions and participated actively at international virtual events (6th International Conference on the Use of Big Data, Mobile Tartu Conference, Asia Pacific Statistics Week 2020 and at the ITU World Telecommunication/ICT Indicators Symposium 2020).

Part II. Ongoing deliverables

- In 2019 the Task Team decided to establish 6 subgroups to work on developing guidelines and methodologies on the use of mobile phone data on specific areas namely: disaster and displacement statistics, dynamic population mapping, information society statistics, migration statistics, tourism statistics, transportation and commuting statistics.
- The subgroups conduct regular meetings in 2020 and 2021 via teleconference to discuss the progress of the preparation of the handbooks. The Wiki is used to share information between members during the drafting process of the respective subgroups. The final draft of the subgroups Handbook is expected to be ready in early 2021.
- The sub-groups are developing handbooks on the use of mobile phone data on specific areas namely: disaster and displacement statistics, dynamic population mapping, information society statistics, migration statistics, tourism statistics, transportation and commuting statistics.

- Members of the task team is preparing a contribution to IOM Practitioners‘ Guide on Harnessing Data Innovation for Migration Policy: Harnessing Data from Mobile Network Operators for Migration Statistics
- Members of the task team are implementing the ITU Big Data pilot projects in Indonesia and Brazil.

Part III. Planned deliverables

- Each sub-group will finalize the handbooks and develop training materials, including e-learning courses. The six handbooks will be published online.
- Development of methods for producing synthetic mobile phone data
- Selected task team members will work with the Task Team on Training, Skills and Capacity Building in preparing the training materials for each of the 6 topic areas.
- Conduct training workshops (both online and face-to-face) to support regional capacity development on the use of mobile phone data in the national statistical offices, in collaboration with regional hubs.
- Members of the Task Team will prepare to participate at international events (7th International Conference on the Use of Big Data, World Data Forum, ISI).

Task Team on Scanner Data

Part I. General Overview

– Members of the Task Team

Chair: Tanya Flower, Office for National Statistics

National Statistical Agencies	Australia Austria Belgium Brazil Canada Denmark Finland Germany Italy Mexico Netherlands New Zealand Norway Poland Switzerland Thailand Turkey United Kingdom United States
International Agencies	Deutsche Bundesbank Eurostat UNSD
Other	Monash University, Australia University of Graz, Austria

– Objective

Increase the use of new data sources (web scraped and scanner data) in consumer price statistics.

– Main outputs so far delivered by the task team:

The task team was relaunched in summer 2020 to build on the success of the 1st phase of the Task Team (May 2017 – April 2019). The deliverables of the 1st phase were:

- Two training workshops on Consumer Price Index calculation were delivered: initially, at the 4th International Conference on Big Data for Official Statistics in Bogota and subsequently, at the 5th International Conference on Big Data for Official Statistics in Kigali.

- The FEWS price index method has been implemented using open source code and is available for use on the UN Global Platform. This implementation also includes the provision of some synthetic data that can be used for testing this method.
- An instructional guide for price index methods that can be applied to these new data sources has been available on the Global Platform since March 2019. It has been peer reviewed by experts from our participating countries.

The 2nd phase of the Task Team is due to run from July 2020 to June 2022. Before the relaunch, a need was identified to expand the existing membership. After inviting expressions of interest, the group expanded to 31 members, covering a wide range of countries and experience. A new term of reference was produced in August 2020. The Task Team objectives are:

1. Price Index Methods - Expand the current offering of methods and documentation relating to the calculation of price indices on the UN Global Platform for Big Data. This will include adding more index methods to the library and reviewing the existing documentation to ensure it remains up to date and relevant
2. Classification methods - Guidance on the process for classifying scanner data to produce data ready for price index compilation. This guidance will include advice on various international approaches to preparing these data including the use of machine learning techniques and, where appropriate, will load methods to the Global Platform for use by NSIs
3. Trusted learning - Ongoing capacity building efforts: expand existing training material; work with the UN Global Platform Team to produce a new certified training course on using these data in consumer prices

Part II. Ongoing deliverables

Price Index Methods

This workstream are currently progressing two deliverables:

1. Work is currently ongoing to update our guidance documentation. The team are currently planning to use a wiki solution instead, hosted on the UNSD wiki site. This provides a more flexible foundation for keeping the documentation as up to date as possible, with more scope to update as the research on this area continues.
2. In parallel, the team are also progressing work to deploy some price index methods onto the UN Global Platform code repository, as well as ensuring there are sufficient data available to test and understand these methods from a training perspective.

Classification methods

This workstream is currently working on developing guidance on the potential scope of classification methods that are available for these data sources, ranging from the simpler methods that can be applied with limited IT requirements, to more complex methods that require data scientist skills and a sufficient IT infrastructure in place.
Once the guidance has been developed (and linked onto the wiki mentioned in the price index methods workstream), the team are then planning to deploy relevant methods/ code onto the UN Global Platform, in a similar manner to the Price Index Methods work.

Trusted learning
This workstream is working in collaboration with the other sub-teams to develop new training content. This will be aligned to the UN Task Team on Training, Skills and Capacity, and their guidance on how to set learning objectives at the different curriculum levels (awareness, beginner, practitioner) and good practice in course design. The team have begun to review existing training content already available from member institutions and will soon start work on scoping out what an awareness level course should incorporate.

Part III. Planned deliverables

We are aiming for these three deliverables – a fully functioning wiki covering all aspects of using these data in consumer price statistics, a documented and tested code repository with methods including classification and price index methods, and an awareness level training course to be ready in time to be presented at the next UNECE Group of Experts on Consumer Price Indices, scheduled to be held virtually in June 2021.

The next phase of the Task Team will then look to continue to expand and update the wiki and code repositories, perhaps focusing on other areas of interest such as product definition. The team will also continue to develop further training materials, at beginner and practitioner level. As the UN Task Team on Training, Skills and Capacity continues to develop, we will align to their guidance in how to begin rolling out these training courses to interested parties.

If interest from other areas, we could also look to expand the use case for these data beyond consumer price statistics, although we would then need to consider reviewing the membership to ensure we have the level of expertise required in these new areas to ensure the documentation and guidance will continue to a sufficient standard.

Task Team on Big Data for the SDGs

Part I. General Overview

– Members of the Task Team

Chair: Niels Ploug, Statistics Denmark

National Statistical Agencies	Country names: Jordan, Poland, Rwanda, United Kingdom, Denmark
International Agencies	Agency names: UNSD, UNESCAP, World Bank, UNICEF, OECD, UNCTAD, FAO
Other	Names of institutes or companies: GPSDD, Bridges 2030, ITU

– Objective

To support implementation at all levels, the 2030 Agenda for Sustainable Development recognizes the need to exploit the contribution to be made by a wide range of data including Earth observations and geospatial information.

The Task Team on big data for the SDGs aims to provide Big Data/non-traditional data tools for a concrete monitoring of SDGs indicators. The ongoing work on Big Data for statistics comprises various aspects of the statistical follow-up, however, at the moment there exist no concretized application of Big Data that can be used by individual countries in the statistical follow-up on the SDGs. There is, however, quite a number of initiatives that can help monitoring SDGs indicators on a global level, including by supplementing the existing data sources. The task team on Big Data for the SDGs is working on an inventory and synthesis of those activities so that they can be applied for monitoring of the SDGs by individual countries.

– Main outputs so far delivered by the task team:

Restart of Task Team activities with a sequence of meetings narrowing the scope of the work. Against this background, the group identified the following streams of work:

- compiling an inventory of examples from countries where big data/non-traditional data has been used for the compilation of SDG indicators;
- bringing transparency to SDG indicators, which offers greater potential for exploiting big data (split up into three categories: earth observations, citizen generated data, and poverty estimations based on Big Data/earth observations);
- advocacy and communication.

- Methodology
The current work is primarily conducted as desk study.

- Projects
Survey within the Global Working Group on big data to identify, which of the 169 SDG targets could be monitored by Big Data/non-traditional data, as well as discussions of big data/non-traditional data sources suited for monitoring specific SDG indicators.

- Events
Participation in various events, such as GWG Big Data Conference.

Part II. Ongoing deliverables

– Methodology

Mainly conducted as desk study and exchanges with relevant stakeholders.

– Projects

Producing an inventory of ongoing big data/non-traditional data activities that can be used for monitoring the SDGs, either as a primary or a supplementary data source.

Creating a data community to provide a platform for exchange of knowledge for application of non-traditional data sources for compilation of the SDG indicators. One of the aims of the platform is to conduct match-making between existing initiatives and countries.

– Events

Side event at the 53 UN Statistical Commission- '*Big Data and the SDGs - what is the way forward - an interactive exchange of views'* Tuesday, 9 March 2021 at 9am New York Time.

In 2021, it is planned to share activities of the Task Team with relevant stakeholders on a regular basis and to update the website of the Task Team.

Part III. Planned deliverables

– Projects

As in Part II. ongoing deliverables

- Training

  In 2021, we plan to share activities of the Task Team with relevant stakeholders on a regular basis and to update the website of the Task Team.

- Events

  Webinars and side events bringing transparency to SDG indicators, and offering greater potential for exploiting big data and a possibility for a knowledge exchange.