UNITED NATIONS

United Nations Statistical Commission

Global Working Group on Big Data for Official Statistics

GWG Bureau meeting

First meeting – 21 March 2018, 13:15 EST

By video-conferencing

Agenda item (3) –

Ongoing projects – Note by UNSD

This additional note describes briefly the status of 3 ongoing projects: (i) UNSD with AWS on Trade Data Lake; (ii) UNEP with Google on Earth engine for measuring Fresh Water Extent; and (iii) FAO with Google Earth on a number of indicators related to SDG 15.

(i) UNSD with AWS on Trade Data Lake

Progress report of the project is given in the Annex below. A few points which can be picked up: AWS gave 24,000 credits (= $24,000.-) for this project with end date 30 June 2019. We have to build up experience to understand how far each credit stretches: What can we store and process? Can we use all necessary tools and application? Etc.

The GWG bureau is requested to give the go ahead, so that UNSD can continue its regular discussions with AWS in developing the Trade Data Lake. Some guidance on connection to common platform, use of credits and reporting back to GWG bureau is appreciated.

(ii) UNEP with Google on earth engine for measuring Fresh Water Extent (SDG 6.6.1)

Here are comments provided by Jilian Campbell of UNEP on the cooperation with Google (and others):

The methodology that we have submitted for SDG 6.6.1 which includes information on how the EO product will be used for SDG reporting is attached.

Here is a bit of a long presentation that I (Jilian) gave for a meeting within my Division which explains quite a bit on the status and approach, it also includes the website that NASA created on our partnership in the presentation: https://prezi.com/view/SyvXqM1zAanw16FHD1yo/

To summarize our approach has been as follows:

1. **Determine the need:** For 6.6.1, we would like the following: (1) water body extent; (2) wetland extent; (3) water quality information for all water bodies; (4) ground water data; (5) water volume data and (6) flow data. For 14.1.1, we are interested in: (1) ocean eutrophication/water quality; (2) marine litter; (3) beach litter; and (4) coastal eutrophication potential (this is based on fertilizer use/run-off, unmanaged waste water, etc.). Ideally, we would like water body extent, wetland extent and water quality as part of a land account approach – ideally we are also interested in land interactions: for example, if a the Juba River basin extent declined can we also see an increase in agricultural production at the upstream end of the basin, can we look to see what land type now exists in places where water once existed.

2. **For each of these topics we have slowly tried to identify partners who can help us fill in the knowledge/data gaps.** To what extent can we use global data. Here are some examples of how this is working so far: For water body extent, there was an existing JRC initiative which had already started working with Google – this is how our partnership started. For wetlands, there is not a global product of high enough quality for the SDGs (CCI-LC is only 300m resolution and thus misses heaps of wetlands, particularly along rivers) this is a research area for ESA/JRC and they are interested in collaborating with us. For water quality, there is work being done by many agencies and NASA is interested in taking a lead in terms of taking this forward (NASA brings with them University partnerships within the USA). For marine litter, there is a global conference on using remote sensing for marine litter in Hawaii coming up, the partners that we have talked to say that the information is not yet there, but many Universities are working on this area.

3. **Convince our partners and countries to pilot test with us:** For the water body data, we included it in our proof of concept in an existing UNEP project last year, we emailed data out to all 193 member states, and we had a high quality pilot with NASA in 9 countries. We are also hosting a workshop for LAC countries on 24-25 April on water body extent. I have a presentation on how Raster data can be converted into tables and land accounting formats which I have been using to explain how maps = data (this is not well understood by most of the countries where I have been working – I have been working with more LDCs and not Mexico/Brazil type of NSOs). See my presentation attached.

4. **Once the data is sound and we have a clear way forward then we have a clear ask to Google**. I would think that Google will also be interested in wetlands and water quality, but we are still trying to define exactly our approach. For water bodies, Google will ensure that we can produce annual data by running the algorithms for the generation of the water quality geospatial data, for aggregating the data at the national, state/province, basin and sub-basin level and by providing the cloud computing. They are also helping us to find ways to make the original geospatial data more accessible to any user (so that water body maps could be downloaded by an 8 year old, not only by a programmer). Google is also hosting the original data, just the volume of data that we currently have (without even running any analysis on that data) would overwhelm our servers.

(iii) Cooperation between FAO and Google on Forests through Collect Earth

The following are comments provided by Marcelo Rezende of FAO.

The main outcome of the partnership in the everyday work of FAO is the bridge they have in Collect Earth with the Google Earth Engine. I am showing below the content of an email I prepared for a conference last year with a summary of the outcomes of the partnership.

For more information on the partnership itself (content of the MoU between FAO and Google) you should contact Danilo Mollicone (focal point for this –danilo.mollicone@fao.org). A concept note on the software itself can be found in **this link** and additional information in the Collect Earth website.

Comments:

I will limit my comments to the application Collect Earth, which continues to be improved and has been used to support many countries in their land monitoring needs.

There are dozens of news on Collect Earth and how the partnership is helping to bridge local knowledge and high end technology. A quick Google search will bring many results. I would like to highlight this one that came out last week: Suite of free, open-source tools to help even non-experts monitor large-scale land use change. The speech of Director-General José Graziano da Silva on the partnership is also very inspiring ("FAO and Google join forces to strengthen the role of technology for sustainable development")

On the SDGs, the publication FAO AND THE SDGs Indicators: Measuring up to the 2030 Agenda for Sustainable Development (page 36) mentions specifically Collect Earth and the Open Foris tools to support countries in monitoring:

- Indicator 15.1.1 Forest area as a percentage of total land area (Tier I);
- Indicator 15.2.1 Progress towards sustainable forest management (Tier II);
- Indicator 15.4.2 Mountain Green Cover Index (Tier II);

Other insightful mentions are:

On SDG 15.3:

Giving new life to degraded lands in Small Island Developing States: An FAO partnership with Google is helping collect and analyze land degradation data

http://www.fao.org/news/story/en/item/1037176/icode/

"For the first time, all countries can equally contribute to global assessments of degraded land and define their own restoration opportunities," said René Castro, Assistant Director-General of FAO's Climate, Biodiversity, Land and Water Department, at COP13 which is being held in Ordos, China."

On SDG 15.4.2:

Mountain Partnership Tracks SDG Indicators, Highlights Regional Climate and Food Security Initiatives

http://sdg.iisd.org/news/mountain-partnership-tracks-sdg-indicators-highlights-regional-climate-and-food-security-initiatives/

'In its report titled, 'Stepping up for mountains: Mountain Partnership Secretariat Annual Report 2016,' the Mountain Partnership reflects on progress as well as challenges in promoting the mountain agenda in 2016, with a focus on climate and food security issues. The report **outlines progress in tracking the Mountain Green Cover Index, one of two indicators for SDG target 15.4, which focuses on the conservation of mountain ecosystems. Data sources for the Index are primarily derived from Collect Earth, a global land use and land cover monitoring project that is jointly implemented by the Food and Agriculture Organization of the UN (FAO) and Google,** as well as the 2015 global map of mountains produced by FAO and the Mountain Partnership Secretariat.'

Perhaps one of the most successful outcomes of the partnership is a publication in Science Magazine on the Extent of Forest in Drylands. It is the **first FAO lead scientific report in Science** and according to Science Altmetrics, it is in the top 5% of all research outputs published by the journal. If there is the possibility to display a video at the exhibit section during the conference, there is a very interesting video demo of Collect Earth.

We have many Collect Earth presentations available. This presentation is about the Science paper and has a nice sequence of slides showing how places that seem empty of trees can host vigorous forests (slide 10 shows a place and then slide 11 zooms in, same for slides 12 and 13). More information on the tool and results from Cape Verde can be found in this presentation. It is more technical but it has many photos and relevant schemes that can be used (credit can be attributed to FAO/Marcelo Rezende).

Statistics on the use Collect Earth can be found in this link. We have now trained over 2000 people in over 40 countries (stats from Feb 2017) and the Collect Earth application/methodology has been published in this peer-reviewed paper.

ANNEX - Project "Trade and Transport Data Lake (T2DL)"

Background

International trade statistics datasets have been available publicly in electronic format (through UN Comtrade) since many years ago. These data are submitted by countries to UN Statistics Division (UNSD) for data harmonization, checking and dissemination. Then users would extract and download the data through UN Comtrade interface or APIs for further processing and analysis. With additional users demand on more real time and disaggregated data (i.e., measurement of cross border e-commerce), traditional international trade statistics datasets are not enough; and must be supplemented by other sources such as UPU parcel post data, ICAO aviation data, or Automated Vessel Identification tracking data. Therefore, the first objective is to establish new data platform to store these diverse datasets, ensure continuous data feed and provide various data analysis and visualization tools. In addition, as a secondary objective, we are exploring the possibility of using the platform to store/disseminate national trade data (possibly disaggregated). Another objective is to host a common trade data processing system based on EUROTRACE.

For more details, see UN Global Platform Proof of Concept Proposal on Data Lake for International Trade and Transport.

Progress Report Q1-2018

The progress made up to 14 March 2018 is as follows:

- Completion of UN GP Proof of Concept Proposal on Data Lake for International Trade and Transport. The proposal contains two objectives, namely 1) Integration of various trade and transport datasets into cloud based data lake environment and 2) Provision of cloud-services to national statistical systems to store, process and disseminate trade statistics.
- Presented the trade data lake project on "PreView" Data workshop organised by German Federal Foreign Office on January 16-17, 2018. PreView is early warning system to gather knowledge on data, tools and implementation techniques from international communities for their crisis management project. The project was well received by the federal office and the fellow speakers/participants.
- UNSD has received and activated complementary AWS credits on March 2018 to be used on various cloud services offered by AWS. The total credit worth is $24,000 with expiration date on 30 June 2019 (duration cannot be extended). There is limitation on which AWS services to be consumed by AWS complementary credits[1].
- UNSD has started to explore the AWS services that are suitable for data lake project. AWS has arranged a meeting with their solution architects to assist us; This meeting would help us avoid "mistakes" in choosing right AWS services.
- UNSD has started migrating UN Comtrade data to Amazon Redshift, a data warehouse solution from AWS.
- UNSD has contacted ICAO; and has received its support to feed aviation data to the data lake. We are currently analysing the data structure; and exploring technical options on storing those data.
- CARICOM expressed its interest to work within scope of the project to store national trade data and to run analysis on the data.

[1] https://aws.amazon.com/research-credits/faq/#services

- UNSD is exploring options on how to use AWS services to run existing version of EUROTRACE software. Several Caribbean countries have expressed interest to use AWS for hosting their EUROTRACE.

Technical Note on Implementation Update

PHASE 1

To set up the minimal data lake services with sample data for testing. The following is the progress on the phase 1:

1. **Data loading:** Sample comtrade data (of 74 million records) is prepared and loaded into **AWS S3** in text format.
2. **Data processing: AWS EMR** (Map reduce service) is used to remove the headers in the files and convert them into parquet format (columnstore format to efficiently integrate with other AWS services).
3. **Data analysis:** Data is loaded into **AWS Redshift** datawarehouse store using COPY command and is ready to do analysis.
4. **Data visualization:** The sample data loaded in redshift is visualized using **AWS Quicksight** to test the performance.

To do:

5. **Data dissemination: AWS lambda function** to be written to expose the s3 and redshift data as RESTful API using **AWS API gateway** service.
6. **Data integration (with other sources):** When data from ICAO is received, set up the **AWS Athena** service to do interactive analysis/visualization with already loaded comtrade data.

General Work Plan 2018-2019

- Continue integrating UN Comtrade, UPU and ICAO datasets to the data lake; and at the same time build and deploy analytical and visualization tools;
- In cooperation with selected regional organisations or countries, design, develop and deploy cloud-based system to store, disseminate and analyse country trade data;
- In consultation with EUROTRACE User Community; explore possibility of upgrading existing EUROTRACE to be cloud-compliant application;
- Integrating the chosen solutions with those of ONS Platform.