Data Access and Data Science

Panel 1(d)

Yogyakarta, 8 November 2022

7th International Conference on Big Data and Data Science

The Fundamental Principles of Official Statistics

Bearing in mind the **critical role of high-quality** official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, **demanding openness and transparency**

Principle 2. To retain **trust** in official statistics, the statistical agencies need to decide according to strictly professional considerations, including **scientific principles and professional ethics**, on the **methods and procedures** for the collection, processing, storage and presentation of statistical data.

Principle 5. Data for statistical purposes may be drawn from **all types of sources**, be they statistical surveys or administrative records. Statistical agencies are to choose the source **with regard to quality, timeliness, costs and the burden** on respondents.

Central Question

If we obtain data from '**outside**',
how much **information/influence/control** do we need
to make sure that we understand its **quality** and
can **explain** how we compile our statistics?
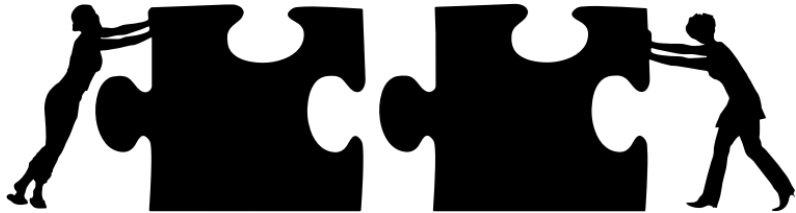
What is quality?

Wikipedia (Garvin, Harvard)
1. Performance
2. Features
3. Reliability
4. Conformance
5. Durability
6. Serviceability
7. Aesthetics
8. Perception

Eurostat (ESS QAF)
1. Relevance
2. Accuracy
3. Reliability
4. Timeliness
5. Punctuality
6. Coherence
7. Comparability
8. Accessibility

We expect you to participate!

Mentimeter:

➤ Go to www.menti.com

➤ Enter the code 11 49 55 2

Audience question 1

Do you believe quality is important for statistics?

1) Never, just give me the numbers!
2) Sometimes, some statistics benefit from quality
3) Often, many statistics need quality indicators
4) Always, statistics without quality are meaningless

Panelists

Daniel Boller, ADB, Manila

Eric Deeben, ONS, Newport

Evest Defiana, Positium, Tartu

Joseph Bulan, ADB, Manila

Muna Al Mazrouei, FCSC, Dubai

Veronika Vilgis, Dataprovider.com, Groningen

Barteld Braaksma, CBS, The Hague (chair)

An Example of Using Satellite Imagery Based-Poverty Data for Targeting:

Bayan-bayanihan Program

EXPLAINER

How Satellite Data Helped Get Food to the Hungry during COVID-19

A COVID-19 emergency food program in the Philippines offered an opportunity to design a targeting program based on granular poverty maps that were compiled using traditional and innovative data sources and artificial intelligence. Photo credit: ADB.

How Satellite Data Helped Get Food to the Hungry during COVID-19 | Development Asia

ADB's Ongoing Work on Poverty Mapping Using Big Data

Japan Fund for Prosperous and Resilient Asia and the Pacific (JFPR) Technical Assistance

- The **Development of New Statistical Resources and Building Capacity in New Data Sources and Technologies** Project aims to equip policy makers with the necessary tools to understand the impact of shocks on different sections of the economy and society in a timely and granular way.
- It explores improving the method of mapping the spatial and temporal distribution of poverty.
- The TA proposes integrating innovative data sources with census and household income surveys, to produce more granular poverty maps for Indonesia and the Maldives.
- Collaborating with World Data Lab and other development partners.

DATA PROVIDER. COM

Access the indexed web

Monthly updates on:

> 600 million domains

> 25 million company websites

> 200 data fields

> 4 years of historical web data

> 5 unique proprietary scores

> 50 countries
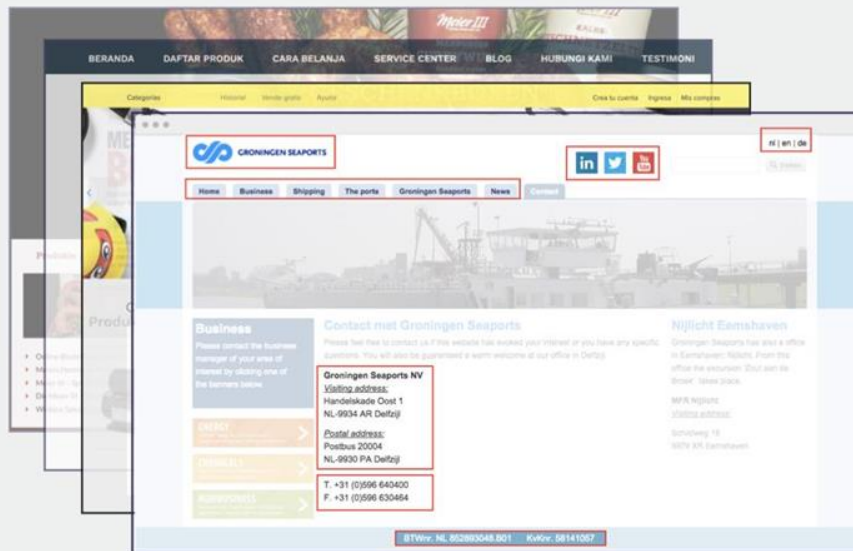
> 200 variables

Geolocation

Business

Content

E-Commerce

Marketing

Technical

Hosting

"A business registry based on the internet." Or "A monthly summary of the internet."
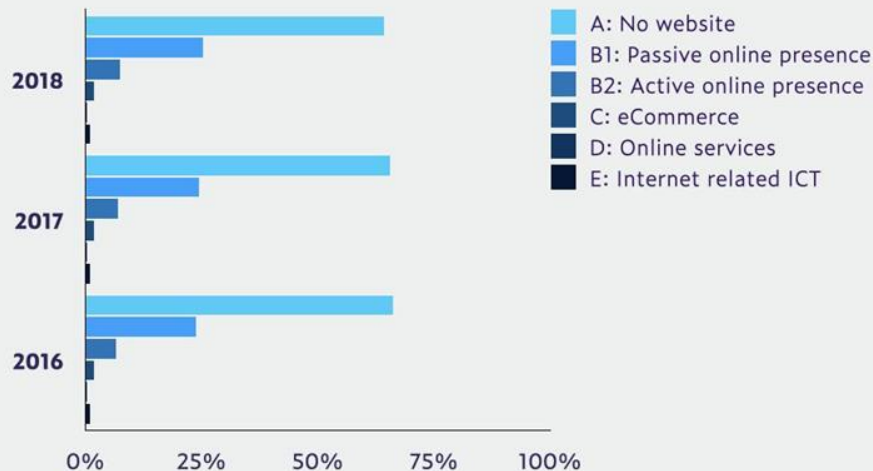
Measuring the Internet Economy

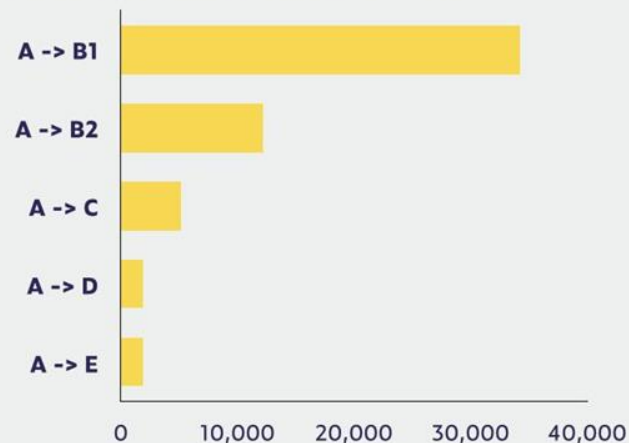Merging web data with national business registries

2016 – 2018: CBS (Netherlands)

2019 – 2021: UN ECLAC with DANE (Colombia), INEGI (Mexico), INE (Chile), IBGE (Brazil)
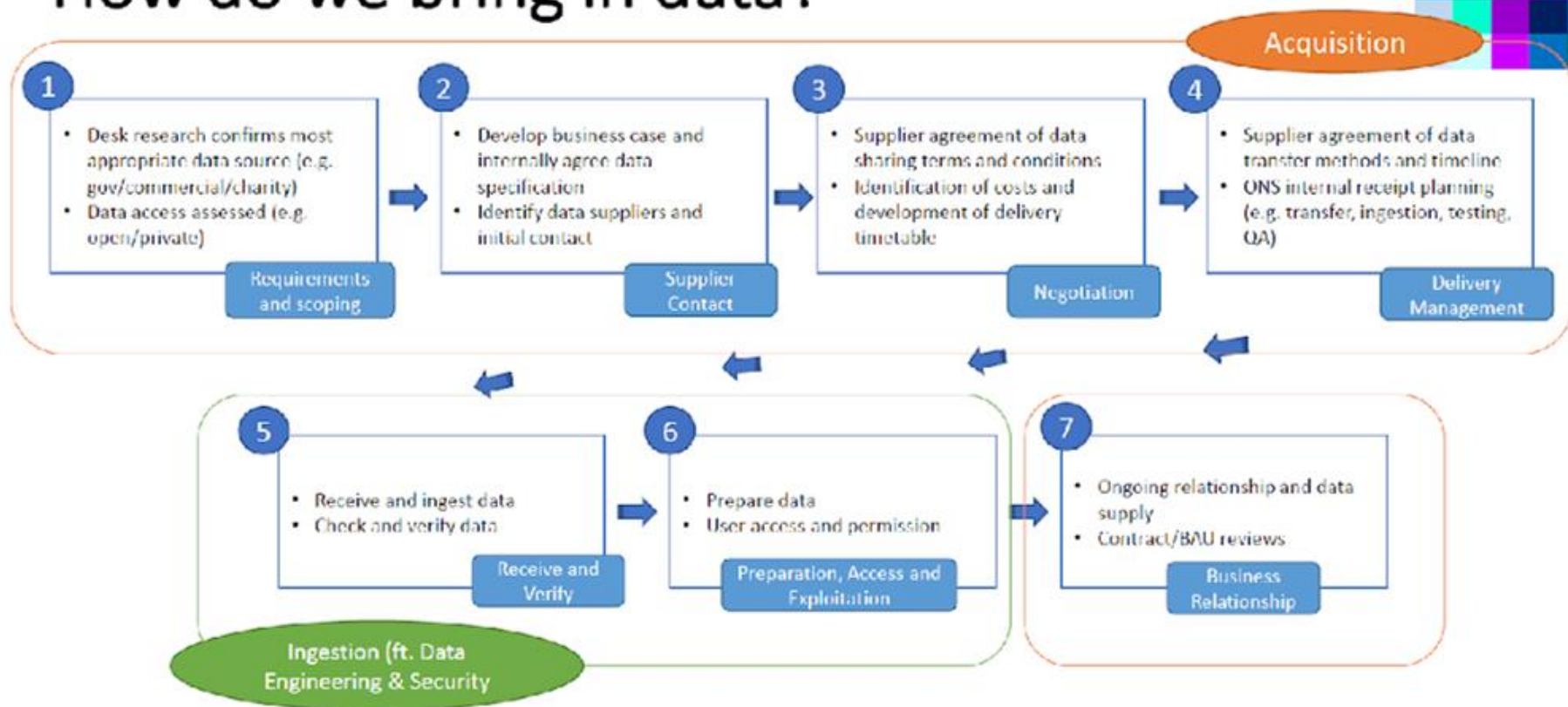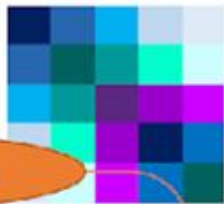
Internet Economy 2016-2018: Netherlands

- A: No website
- B1: Passive online presence
- B2: Active online presence
- C: eCommerce
- D: Online services
- E: Internet related ICT

2018

2017

2016

0%   25%   50%   75%   100%

Change in Web Presence between March – September 2020: Colombia

A -> B1

A -> B2

A -> C

A -> D

A -> E

0   10,000   20,000   30,000   40,000

Sources: CBS, IBGE, UN ECLAC

How do we bring in data?

Acquisition

1
- Desk research confirms most appropriate data source (e.g. gov/commercial/charity)
- Data access assessed (e.g. open/private)

Requirements and scoping

2
- Develop business case and internally agree data specification
- Identify data suppliers and initial contact

Supplier Contact

3
- Supplier agreement of data sharing terms and conditions
- Identification of costs and development of delivery timetable

Negotiation

4
- Supplier agreement of data transfer methods and timeline
- ONS internal receipt planning (e.g. transfer, ingestion, testing, QA)

Delivery Management

5
- Receive and ingest data
- Check and verify data

Receive and Verify

6
- Prepare data
- User access and permission

Preparation, Access and Exploitation

7
- Ongoing relationship and data supply
- Contract/BAU reviews

Business Relationship

Ingestion (ft. Data Engineering & Security

Case Study – Mobile Phone Data

- March 2020: Urgent need to understand how well non-pharmaceutical interventions – such as mobility restrictions – were being observed

- DSC worked with O2 Motion to explore the country's response to lockdown

- Access to anonymised and aggregated O2 Motion telecoms mobility data - 25 million mobile phone customers have helped us to understand mobility trends in the UK population

- Close-to-real-time information - mobility patterns by age, gender, and whether trips were for commuting or other purposes

- Ethical review carried out

- Underlying data and analysis remain confidential - commercial restrictions

https://datasciencecampus.ons.gov.uk/understanding-mobility-during-the-covid-19-pandemic/ and
https://www.o2.co.uk/business/why-o2/customer-stories/o2-motion-and-ons

Data Science Campus
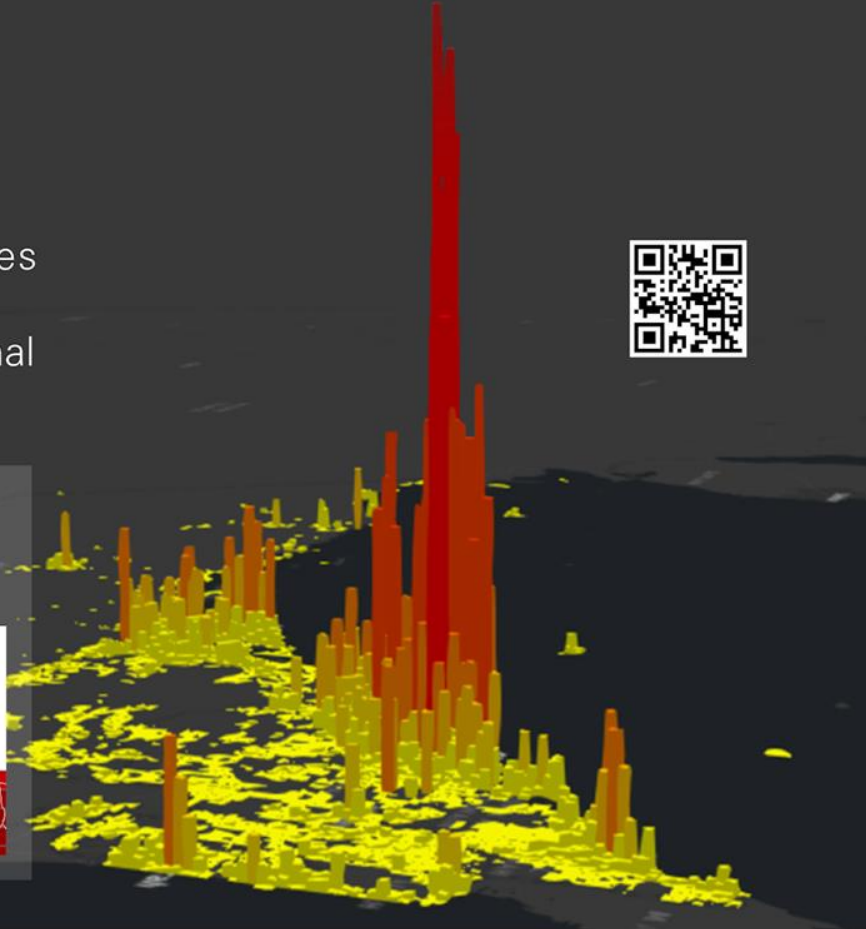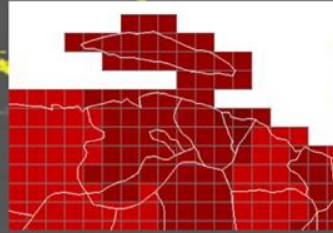
A snapshot of our recent research…and challenges for data quality!

By Stanley Lawrence Sie, Charibeth Cheng, Cedric Basuel, Alyssa Villanueva, Kadra J. Saeed, and Daniel Boller

ADB

Research explores how natural language processing (NLP) techniques can be applied to social media text data to map public sentiment and inform development research and policy making, supported by two case studies related to the coronavirus disease (COVID-19) in the Philippines and the public debate on climate change in Australia.
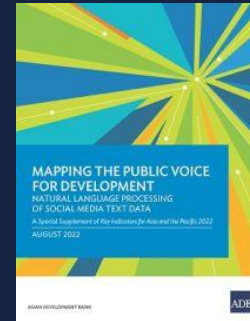
Mapping the Public Voice for Development—Natural Language Processing of Social Media Text Data: A Special Supplement of Key Indicators for Asia and the Pacific 2022 (https://www.adb.org/publications/mapping-public-voice-development)

Case studies showcase the applicability of NLP of social media text data (here: Tweets from Twitter) 🡪
Challenges:
- Only a minor fraction of the population is active on (and uses) social media channels ("social media bubbles").
- Social media users are not (necessarily) representative for the society (and information are often missing).
- ….and many more.

"Although proponents and skeptics alike tend to view the big data revolution as totalizing and universal, the reality is that billions of people remain on its margins because they do not routinely engage in activities that big data and advanced analytics are designed to capture."

(J. Lerman)

Let us discuss opportunities to improve data quality (and data impact)!

Transparency <-> Influence <-> Control
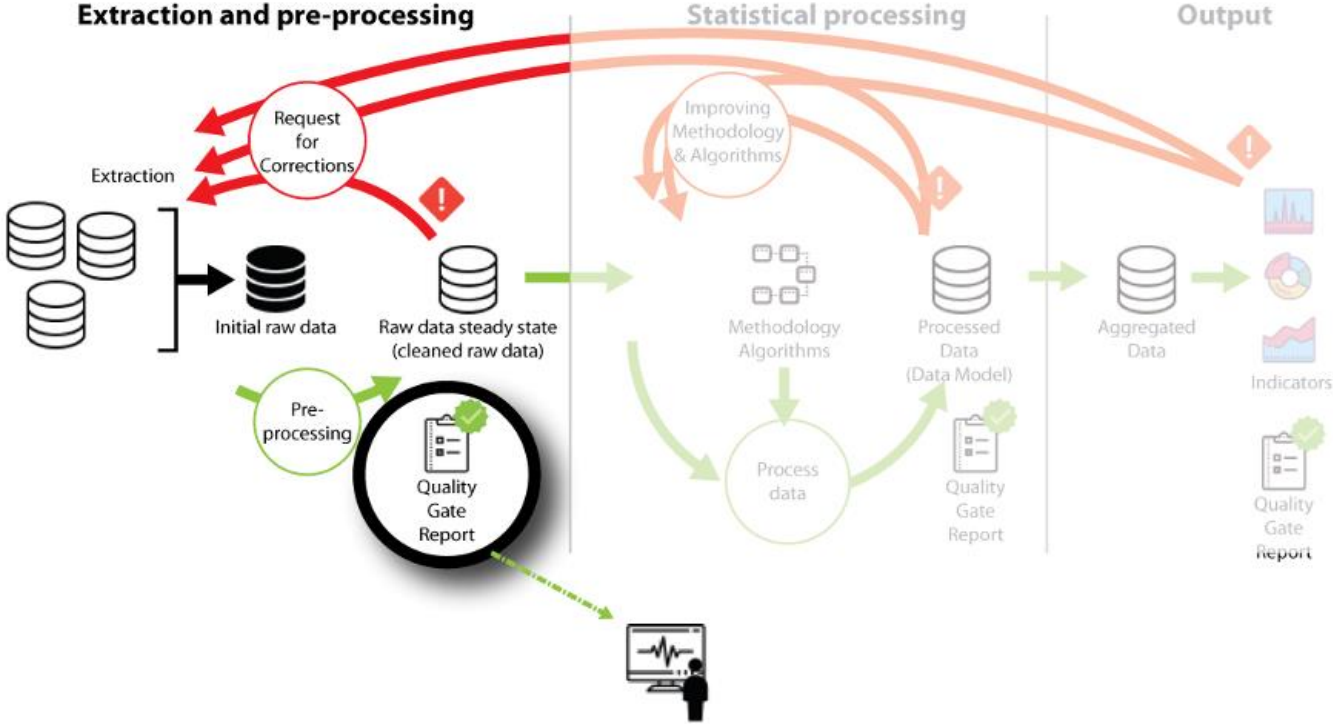
Panel and Mentimeter Questions

Mentimeter:
➢ Go to www.menti.com
➢ Enter the code 11 49 552

Audience question 2

What is your favorite quality dimension?

Open question (word cloud),
multiple entries (max 5) allowed

Panel Questions

- What considerations for data quality are decisive for collaborations between the private and public sector?

- How could satellite imagery be used to elevate national data quality, and what mechanisms do we need to ensure quality?

Audience question 3

What are the key challenges in QA when acquiring data to produce rapid indicators through data science?

Open question (word cloud),
multiple entries (max 5) allowed

Panel Questions

- What considerations for data quality are decisive for collaborations between the private and public sector?

- How could satellite imagery be used to elevate national data quality, and what mechanisms do we need to ensure quality?

- What are the key challenges in quality assurance when acquiring data to produce rapid indicators/statistics through data science; and how do you tackle them?

- How can we all contribute to capacity building with the use of non-traditional data sources as well as improvement in the output quality of the resulting statistics?

Audience question 4

As more and more data is being owned, processed and analyzed by the private sector, is it time to rethink the role of NSOs?

1) Of course, the world is changing and we should too
2) Probably, but let's wait until we are asked to
3) Maybe, but first we have to do research for ten years
4) No way, I never heard complaints about our statistics

Panel Questions

- What considerations for data quality are decisive for collaborations between the private and public sector?

- How could satellite imagery be used to elevate national data quality, and what mechanisms do we need to ensure quality?

- What are the key challenges in quality assurance when acquiring data to produce rapid indicators/statistics through data science; and how do you tackle them?

- How can we all contribute to capacity building with the use of non-traditional data sources as well as improvement in the output quality of the resulting statistics?

- As more and more data is being owned, processed and analyzed by the private sector, is it time to rethink the role of NSOs?

Audience question 5

As more and more data is being owned, processed and analyzed by the private sector, is it time to rethink the role of NSOs?

1) Of course, the world is changing and we should too
2) Probably, but let's wait until we are asked to
3) Maybe, but first we have to do research for ten years
4) No way, I never heard complaints about our statistics