

# Big Data Quality Framework

**Ronald Jansen**

Assistant Director

Chief of Data Innovation and Capacity Branch  
United Nations Statistics Division

# Overview

## Big Data Quality Framework

Input

Throughput

Output

Examples

Change Management

## Big Data Quality Framework

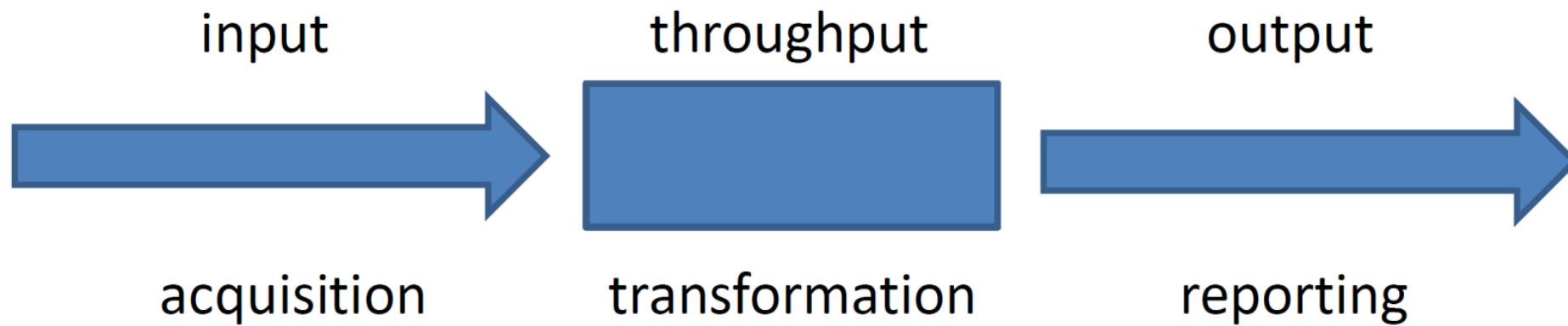
CES

UNECE Task Force

1994

# Approach

## Business process

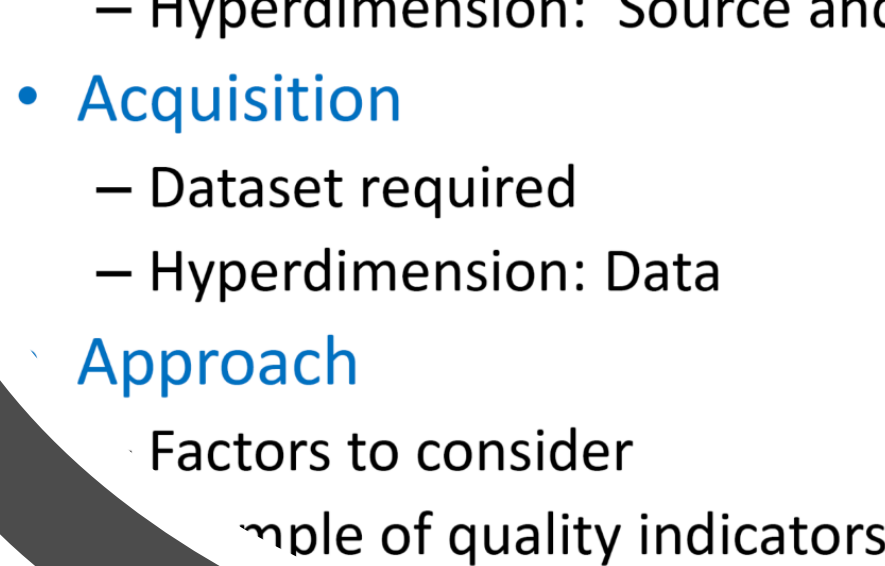



## Framework

Structured view of quality for each phase



# Input

- Discovery stage
    - Dataset not required
    - Hyperdimension: Source and Metadata
  - Acquisition
    - Dataset required
    - Hyperdimension: Data
  - Approach
    - Factors to consider
    - Sample of quality indicators
- 
- 

Hyperdi- mension	Quality Dimension	Factors to consider
---------------------	----------------------	---------------------

## SOURCE

### Institutional environment

Sustainability of the data provider

Reliability of data provider

Transparency of data provider

### Privacy and Security

Legislation

Data Keeper vs. Data Provider

Restrictions

Perception

<b>Metadata</b>	Complexity	Technical constraints, Structured or Unstructured Readability, Presence of hierarchies and nesting
	Completeness	Metadata is available, interpretable and complete
	Usability	Resources required to import and analyse Risk analysis
	Time-related	Timeliness, Periodicity, Changes through time
	Linkability	Presence and quality of linking variables
	Coherence	Use of standards
	Validity	Transparency of methods and processes Soundness of methods and processes

Data	Accuracy and selectivity	Total survey error approach Reference datasets Selectivity
	Linkability	Quality of linking variables
	Coherence - consistency	Coherence between metadata description and observed data values
	Validity	Coherence between processes and methods and observed data values

# Throughput

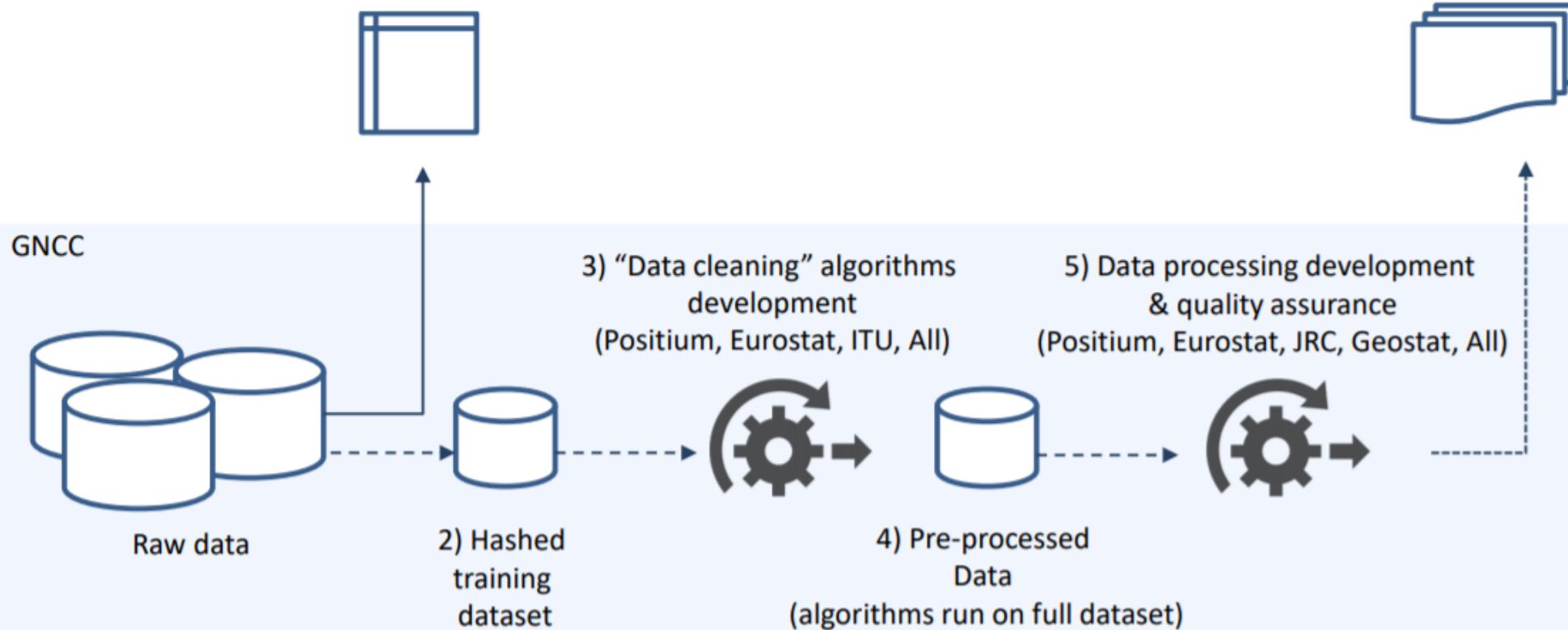
- **System Independence:** The result of processing the data should be independent of the hardware and software systems used to process it;
- **Steady states:** that the data be processed through a series of stable versions that can be referenced by future processes and by multiple parts of the organisation ;
- **Application of Quality Gates:** that the NSO employ quality gates as a quality control business process.



1) Small data sample to understanding data fields and attributes  
(Positium, Eurostat, ITU, JRC, All)

6) Derived products:

- Migration/tourism statistics (Geostat, SCM, All)
- Grid maps (data challenge, JRC, All)



# Output

- Output quality framework should
  - Meet reporting criteria of the NSO
  - Provide required information to allow users to make informed decision regarding the use of the statistical output
  - Follow transparency principle
  - Follow the general approach with quality dimensions, indicators and factors to consider

UNITED NATIONS  
NATIONAL QUALITY ASSURANCE FRAMEWORKS  
MANUAL FOR OFFICIAL STATISTICS  
(UN NQAF MANUAL)  
Including recommendations, the framework and  
implementation guidance

Chapter 3. The United Nations National Quality Assurance Framework: principles and requirements.....	17
Introduction .....	17
Level A. Managing the statistical system.....	20
Level B. Managing the institutional environment.....	22
Level C. Managing statistical processes.....	25
Level D. Managing statistical outputs.....	27

# Examples

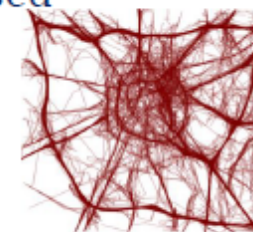
## Quality framework in practice: Selectivity

**GPS data** (Pratesi et. al., 2015): **inhabitants' mobility using data of private vehicles tracked with a GPS device.** The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points. Vehicle traces were then mapped on the road network.

### **Mobility collected on a self-selected sample of car journeys**

Methodological Framework:

- reconcile data from the two independent sources (Big Data and sample surveys) ;
- if no common variables are available, known correlated data could be used;
- use Big Data directly as a covariate under a model-based approach.



# Examples

## Quality framework in practice: Missing data

**Remote sensing for Agricultural Statistics** (Tam and Clarke, 2015b): investigate the use of satellite sensor data for the production of agricultural statistics such as land use, crop type and crop yield. **The coverage of satellite data is the same as the coverage of land parcels.**



**Missing data due to persistent cloud cover.**

Methodological Framework:

- Set up a model where ignorability conditions hold:  
Calibration or imputation techniques;
- Model with Ignorability conditions not found for certain areas ( Weather may affect the type of crops being grown, or yields, missing data related to the target variable).



Use traditional data collections for these areas

# Change Management

Making changes – Maintaining quality

## Example

Statistics Canada

Improving Agricultural crop  
statistics

Use of Surveys

Use of Satellite data

Just Month September

Reduce response burden

Maintain quality

## Example

Australia Bureau of Statistics

Calculating CPI

Use of Price Surveys

Use of Scanner data

Replace 25% of Basket

Reduce cost

Gain more detail

Maintain quality

# Change Management

Making changes – Maintaining quality

## Tourism Statistics

### Border Surveys

Positives

Controlled sample

Targeted questions

Negatives

Non-response

Long processing time

Aggregated data

### Mobile Phone Data

Positives

Detail

Short processing time

Negatives

Selectivity

No info on purpose

# Thank you