Quality assurance of official statistics based on privately held data: the use of reference methodological pipelines

Peter Struijs and Fabio Ricciato

Eurostat, Luxembourg

Revision of Regulation (EC) 223/2009 on European Statistics

- Adopted by European Parliament in 03/24
  - https://www.europarl.europa.eu/doceo/document/TA-9-2024-0152_EN.pdf

- Introduces legal enablers, obligations and safeguards, for the reuse of Privately Held Data (PHD) in official statistics

- **Article 17b – Obligation** of private data holders to make data available for developing, producing and disseminating European statistics

- **Art. 17c – Requests** for data and arrangements for making data available for development, production and dissemination of European statistics to be made by NSI or Eurostat
  - Request shall specify: **what data and metadata** are required; the statistical need [...]; the frequency [...] and the deadlines; the **operational arrangements** for making the data available.
  - Following a request for data [..] a **dialogue** shall take place between the NSI [...] or the Commission (Eurostat) and the data holder concerned to discuss and agree on the measures needed for making data available for the development, production and dissemination of European statistics, with the aim of concluding an **agreement**

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Reusing PHD for OS brings new challenges for statistical methodology and quality assurance …

- Raw PHD are typically "digital crumbs" produced as by-product of business services, not designed specifically for statistical purposes

- Typically very granular (e.g., individual transactions), high volumes

- → heavy processing is required to transform "raw" PHD into statistical indicators

- → processing tends to be "heavy": **methodologies are complex to design** and resource-intensive to execute

- → part of the processing may take place at the premises of the data holder

- → comes with multiple new quality assurance challenges

... and requires the definition of a new partnership model
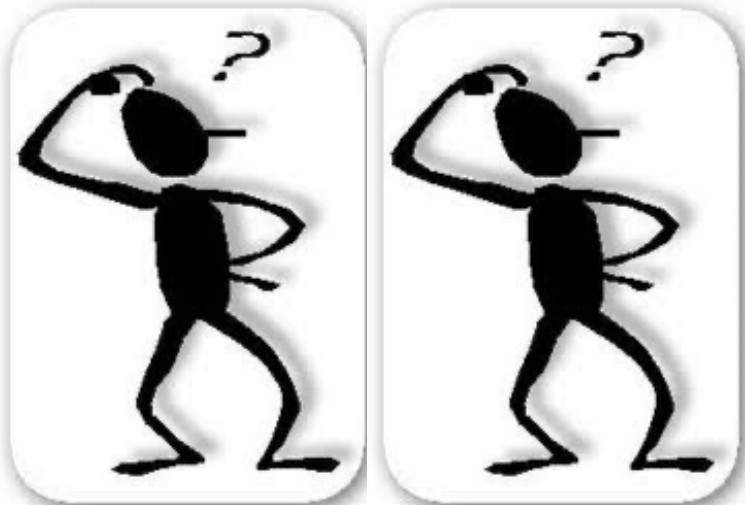
- What the data holder is expected to do? What and How data should be "prepared"? What and How the prepared data are passed to the NSI?

- What meta-data or para-data are to be provided? How the interaction between the data holder and the NSI is to be organized? What should be reported, when, in which form?

- The responsibility for the final statistics is with the NSI, but what are the data holder's responsibilities?

- Financial compensation for data preparation?

- Non-financial incentives?

- What data protection measures should be adopted?

- What business secrets must be protected and how?

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
Statistics Portugal

eurostat

The conference is partly financed by the European Union

Taking inspiration from MNO data: the Reference Methodological Pipeline (RMP)

eurostat

Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System

2023 edition

STATISTICAL REPORTS

- Mobile Network Operator (MNO) data is a particular kind of PHD

- Past ESS work on MNO data led to the concept of **Reference Methodological Pipeline (RMP)** based on a common standardized open methodology framework for the whole ESS

- Formally adopted by the Task Force on MNO data, see position paper[*], and currently under development in two ESS projects

- For MNO data, RMP is a central instrument for assuring "quality"

- Can we apply the RMP concept to other PHD?
  - E.g. Financial Transaction Data (FTD), Smart Meter data, online platforms ...

(*) Position paper by TF MNO available at: https://ec.europa.eu/eurostat/documents/7870049/17468840/KS-FT-23-001-EN-N.pdf
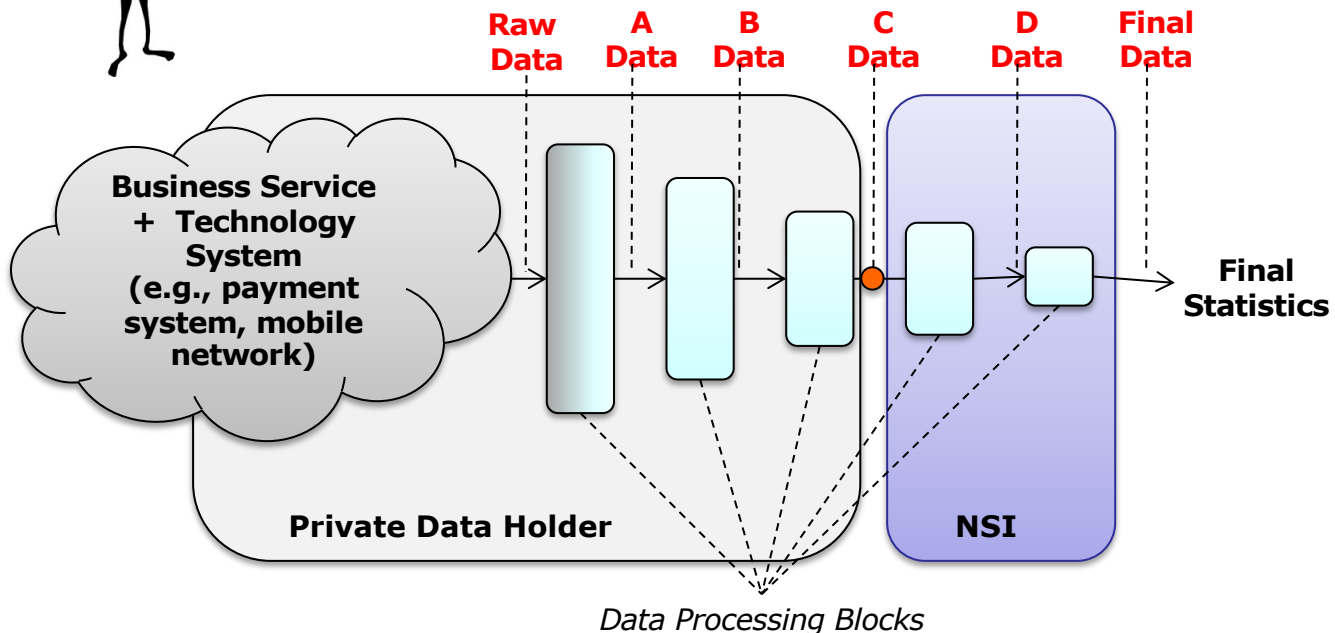
What is a Reference Methodological Pipeline (RMP)
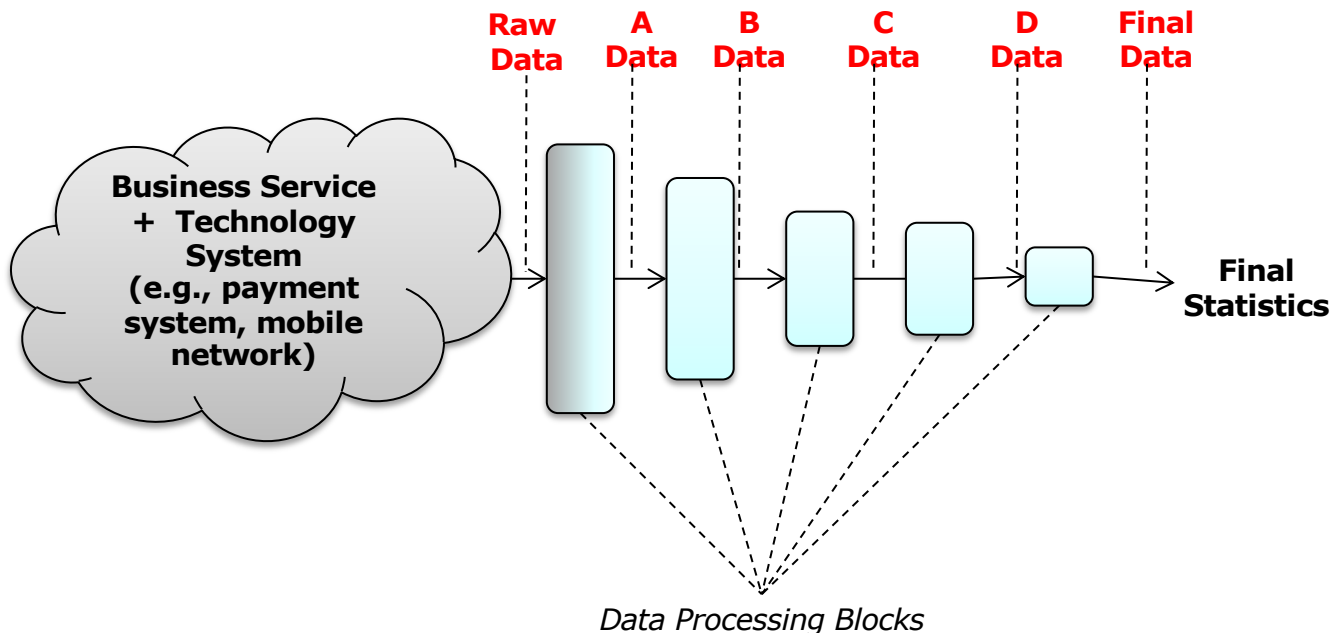
- RMP is a **detailed** and **unambiguous** specification of the whole **end-to-end** data processing flow, from "raw" data to final indicators

- RMP is based a **modular reference architecture** (blocks + interfaces)...

- describing the function (what it does) and the method (how it does) of each processing module, and the semantic of the data objects

- Unambiguous specification → expressed in a formal language

Raw Data · A Data · B Data · C Data · D Data · Final Data

Business Service + Technology System (e.g., payment system, mobile network)

Final Statistics

Data Processing Blocks

Benefits of RMP 1/2

- **Enables the NSI to take responsibility for the end product**

- Represents a complete, open and transparent methodological specification

- Provides a basis to specify quality assurance system

- Enables dialogue between NSI and data holder as to how the *processing execution* is to be split

- Provides a basis to assess and distribute costs and compensations

Benefits of RMP 2/2
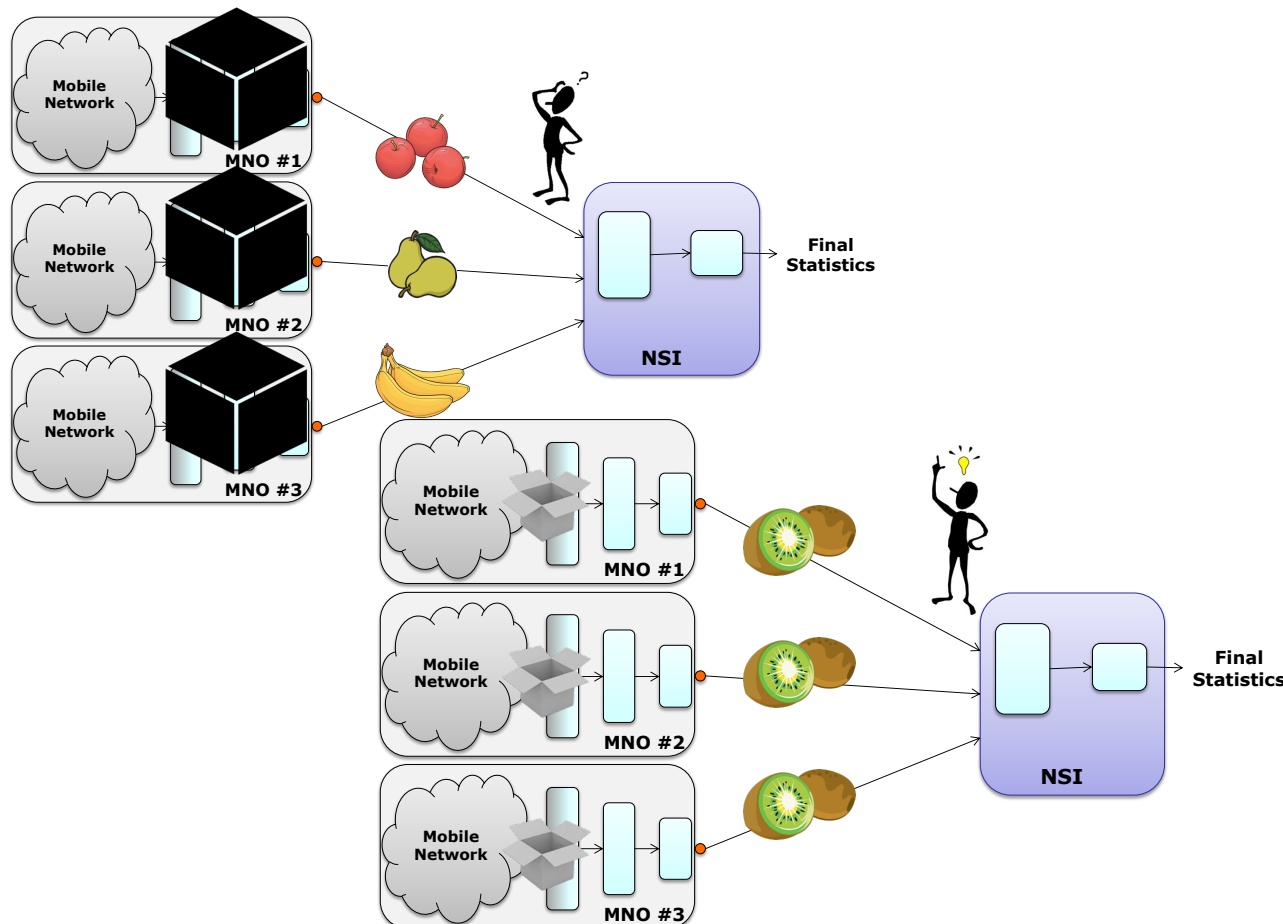
- Ensures comparability and *combinability* of (intermediate) data from multiple data providers

- Provides a basis to assess the confidentiality risk at any stage of the workflow …

- … and identify adequate supplementary Technical and Organisational Measures for compliance with data protection regulation, in coordination with DPAs (e.g., PET)

DPA: Data Protection Authority
PET: Privacy-Enhancing Technologies

Requirements for RMP
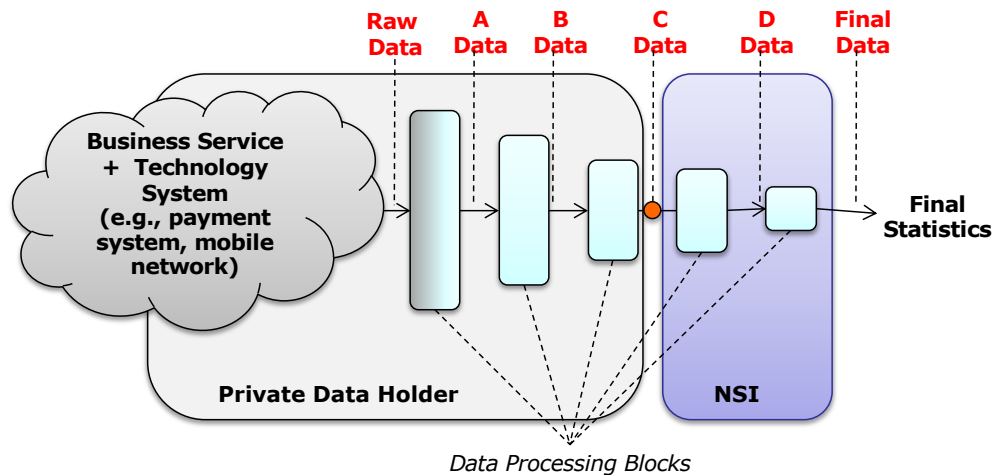
- Modular and Open

- Sufficiently flexible and customisable to accommodate differences across data providers

- Designed to evolve (methods and data objects will change, the underlying pipeline should accommodate such evolutions with incremental additions)

- Accompanied by an open-source software **reference implementation** expressed in a formal programming language(s) + reference test data
  - to enable verification of functional compliance by alternative, possibly proprietary implementations

- NB: the ESS needs to define a RMP governance framework to decide and manage new RMP releases
  - part of future COP/QAF versions?

To RMP or not to RMP?

- ESS RMP requires resources and costs (at ESS level) to be **developed** and **maintained**

- Decision as to whether to adopt RMP or not depends on cost-benefit assessment

- The cost-benefit balance depends on specific type of data source

- Dimensions to consider include
  - Complexity of methodology (data source and use-cases)
  - Amount of processing required at data holder's premises
  - Need of combining data from multiple providers
  - Need of combining data across different countries
  - …

Take-home messages

- Exploitation of new data sources raises **quality assurance** issues that cannot be addressed in the traditional way.

- A novel approach to quality assurance based on RMP is under development in the ESS for MNO data:
the RMP approach *could* be adopted also for other types of data source

- Building upon the current experience with MNO data, the suitability of RMP for other data sources should be based on a careful cost-vs-benefit analysis

- Such an assessment could potentially conclude that RMP is to be taken as the *default* approach to quality assurance with PHD for European statistics

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Thank you!