



# UN Task Team on Scanner Data – 2025 update

Tanya Flower, Federico Polidoro, Claude Lamboray

*Scanner Data Steering Group*

*(Serge Goussev, Collin Brown, Vladimir Goncalves Miranda,  
Lincoln Teixeira da Silva, Helen Sands, Tanya and Federico)*

ICP Inter-Agency Coordination Group (IACG) Meeting

20 - 24 October 2025

World Bank HQ, Washington, DC

# Outline

- UN Committee of Experts on Big Data and Data Science for Official Statistics
- UN Task Team on Scanner Data
- Refreshed workstreams and their main activities.
- Medium term planning
- Other guidance and additional materials still in the works
- UNTT governance - Secretariat responsibilities

# UN Committee of Experts on Big Data and Data Science for Official Statistics



- ▶ Previously the Global Working Group on Big Data for Official Statistics
- ▶ Mandated to give direction to the use of Big Data for Official Statistics

# UN Task Team on Scanner Data



Refreshed aim (2025): enable and expand the practical usage of **alternative data sources** (including scanner) in consumer price statistics. Future work will look to develop the wider usage of these data (for example, in household expenditure statistics).

Task Team [homepage](#)

# Refreshed workstreams



Classification

Initial set of guidance released [here](#)



Training

Initial courses released [here](#)



System architecture

[Report of the survey conducted](#)



Reproducibility/FAIR

Interim guidance released [here](#)



Handbook

Available [here](#)

# Classification (1/2)

**Team:** Serge Goussev (Lead), David Chiumera, Florin Barb, Frances Krsinich, Jens Mehrhoff, Julio Cesar de Azevedo Vieira, Karola Henn, Kjersti Nyborg Hov, Liam Greenhough, Roman Höhn

- ▶ Purpose of workstream is to draft guidance and supporting material on classifying alternative data to prepare it for price index calculation.
- ▶ Classification is a critical task of assigning each price relative to the strata used as input for price index calculation. Critically, this is assigning a product category (such as COICOP category) to the unique product in alternative data. However, it could also include geography-based classification and other methods applicable to the price index method.
- ▶ Most of the guidance that was pre-released via succinct presentations at [CPI EG \(2023\)](#) and [Ottawa Group \(2024\)](#) has been documented in detail and is now published as wiki pages on the e-handbook for everyone!
- ▶ Other guidance and additional materials still in the works

# Classification (2/2)

## Completed and published

- Pre-conditions and deciding on appropriate classification methods
  - Note on variables applicable for classification
- Method 0: Manual labelling or validation of predicted labels
- Method 1: Attribute-based classification method
- Method 2: Pattern matching classification method
- Method 3: Recommendation / Machine-assisted classification
- Method 4: Machine Learning classification method
  - How to evaluate classification methods
  - Working with class imbalance
- Operational best practices
  - Designing the classification step: operational considerations

[Section in e-Handbook here](#)

## Work in progress

Guidance on:

- Blending classification methods
- Dealing with taxonomy changes
- Designing the classification step: Technical considerations
- Other examples and best practices

Other materials:

- Example code notebooks to showcase the 4 main methods
- RAP template that demonstrates how to operationalize classification

# Training (1/2)

**Team:** F. Polidoro (Lead), S. Goussev, K. Nieminen, C. Bontemps, T. Flower, V. G. Miranda, L. T. da Silva, O. Eugster, L. Palumbo; C. Nicholls; V. M. Guerreiro; S. Andric

- ▶ The mission of WS3: develop new training packages using the guidance material to promote the use of these new data sources and methods
- ▶ Courses placed onto the UN Learning Management System (UN LMS) which is held on the UN Global Platform.
- ▶ Delivery style: e-learning through Automated PowerPoint with voice over, short video (in some cases) or guidance sheet (informative). Guided hands-on experience in R and Python when needed and useful
- ▶ The curriculum was updated to adhere to the outline of the e-handbook, and it was presented at the Ottawa group meeting (it is available [here](#))



# Training (2/2)

- ▶ Slowdowns from the last presentation at the Ottawa Group due to:
  - ❑ Very time-consuming process to record the voiceover
  - ❑ The need to revise the materials drafted
  - ❑ Issues around the Learning Platform
- ▶ Advancements achieved:
  - ❑ AI tool adopted to record the voiceover (Speechify; a license acquired by WB for TT use)
  - ❑ The learning platform resumed operations
  - ❑ The first course “Alternative Data Sources (ADS) to compile CPI: an overview” is available online to the NSOs and users and “Data acquisition” will be soon
  - ❑ Course 6 (“Price index methods”) almost finalized
  - ❑ A governance structure has been setup and the WS3 can upload the materials on the UN learning platform on their own
  - ❑ Github repository created to host supplementary material for the course (code, data, and further resources to gain a deeper understanding of ADS for price statistics).
- ▶ More details about the state of play of the work on the training materials available in the poster presented in the poster session

# Reproducibility/FAIR (1 / 6)



Team: Serge Goussev (Lead), Ben Hillman, Caroline White, Chihiro Shimizu, Christophe Bontemps, Claude Lamboray, Federico Polidoro, Jens Mehrhoff, Joni Karanka, Luis Gerardo Gonzalez Morales, Rafael Posse, Sean Lovell, Steve Martin, Tanya Flower

See [about the team](#) for fuller overview

- ▶ In the price statistics discipline, it is very hard to use open data and make our research projects reproducible. As a result, most research projects are very hard to reproduce, hard to expand, and hard to use for training/capacity building.
- ▶ A new workstream was formed to tackle this challenge.
- ▶ What the project does:
  - Provide clear and approachable guidance on how researchers can make their projects reproducible
  - Support cataloguing open datasets so that researchers can know and use them for their research (or capacity building/training).
- ▶ What the project does not:
  - Produce new standards/guidance - instead, focuses on providing an accessible link to guidance already made by other groups such as [the Turing Way](#), [RAP](#), and [FAIR](#).
- ▶ What is published:
  - *Interim guidance at this point!*

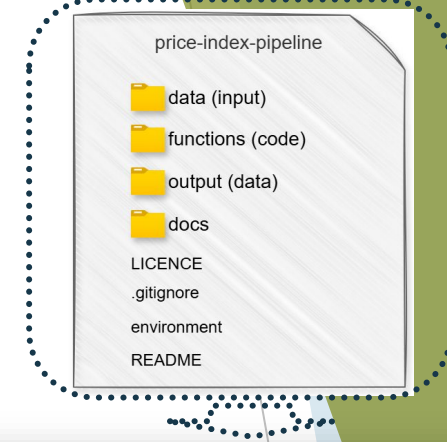
# Reproducibility/FAIR (2/6)

- ▶ A common quadrant can be used to contextualize the differences between reproducible, replicable, robust, and generalizable research.
- ▶ The typical result is generalizability in the projects we work on.
- ▶ The goal of the project is to help researchers make their projects:
  - Goal: Reproducible
  - Backup: Replicable

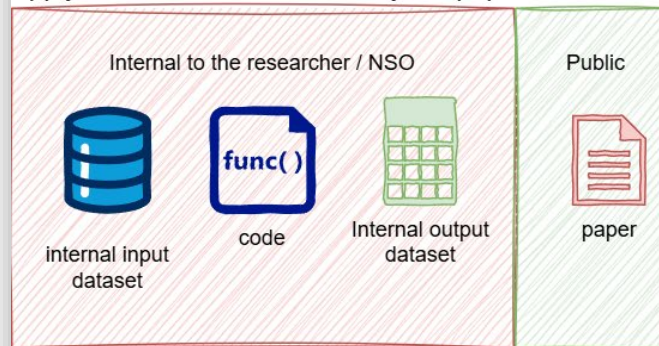
		 Data	
		Same	Different
 Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

# Reproducibility/FAIR (3/6)

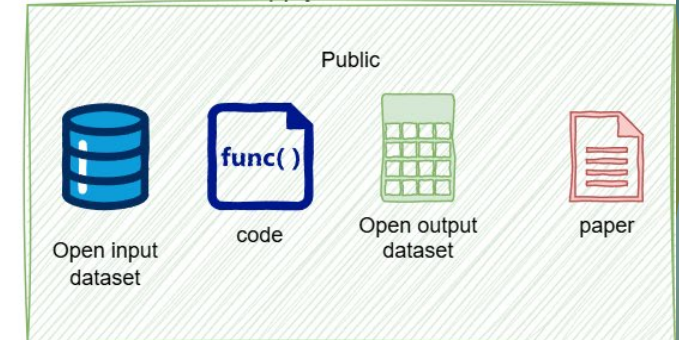
- ▶ Visually this can be shown as follows:
- ▶ If a proprietary dataset is used but the code and example output are made available, then researchers can expand on the findings and others can learn from the research team!
- ▶ Guidance on how to structure reproducible or reusable projects also available!
- ▶ And many more topics!



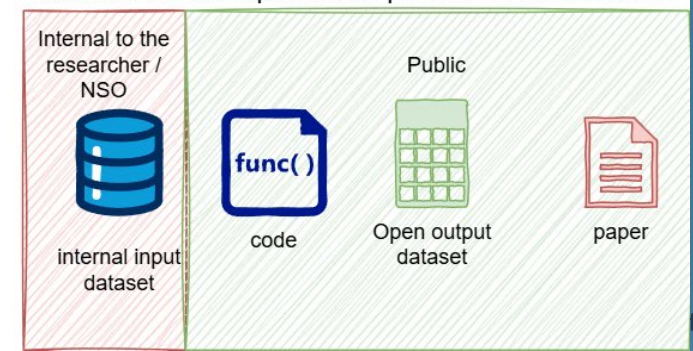
**Default: Generalizability.** Researchers struggle to understand and repeat aspects of the research and apply with their own data as only the paper is available.



**Goal: Reproducibility.** Open data and code is used. Researchers replicate the results or expand them, and can apply them on other data

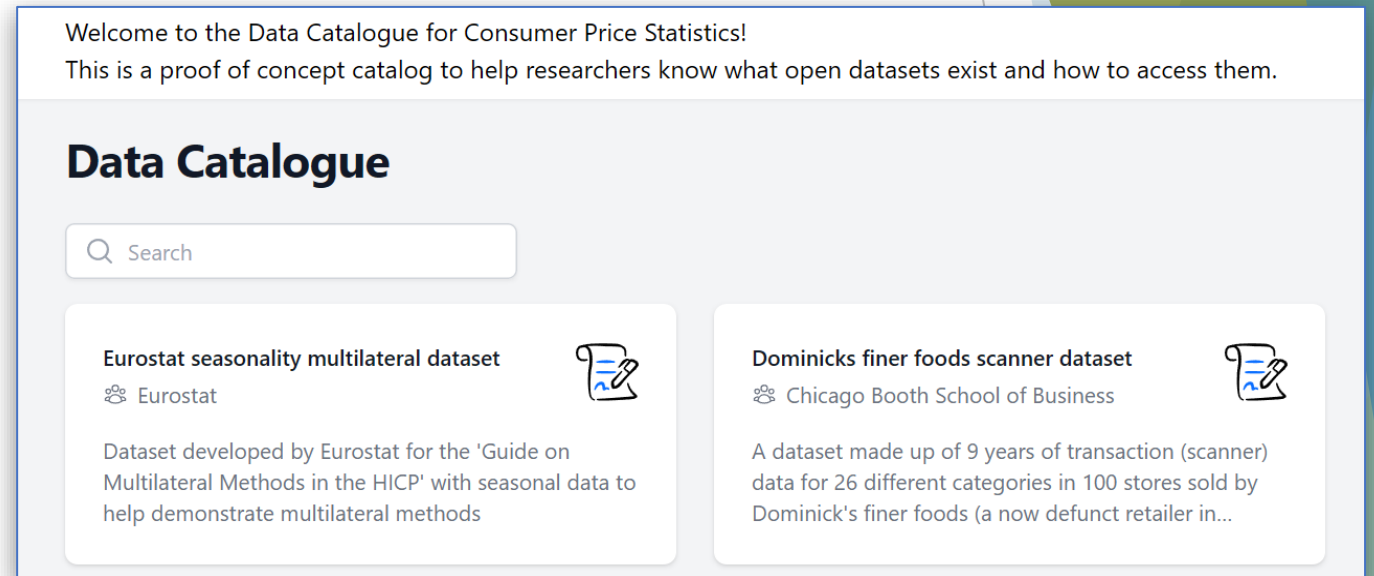


**Backup: Replicability.** Data is not available to other researchers but code is published. Researchers can understand and replicate the process with other data.



# Reproducibility/FAIR (5/6)

- ▶ A basic Proof-of-Concept data catalogue is made available: [Price Statistics Data Catalogue](#)
- ▶ See a dataset missing? Check out the [CONTRIBUTING.md](#) to figure out how to help add more datasets!
- ▶ See [how the catalogue works](#) to understand more about it!



# Reproducibility/FAIR (6/6)

- ▶ Register more datasets to the catalogue (we need your help!)
- ▶ Next steps:
  - Provide more flushed out guidance on reproducibility, including developing maturity levels to simplify onboarding for researchers
  - Make a connection with the FAIR community and integrate the FAIR principles into our guidance
  - Provide guidance on metadata standards for research purposes
  - Provide guidance on synthetic datasets
  - Provide guidance on harder questions - like what to do with widely used proprietary datasets and the catalogue
  - Coordinate with organizational and international bodies (CPI EG, Ottawa Group, IWGPS, etc)
  - And other topics!



# Medium term planning

- ▶ The deliverables listed here are medium priority for the Task Team - it is unlikely we can take these forward until some of the other workstreams complete unless we can identify some additional capacity
  - Discussion forum - we would like to identify a way to enable communication between practitioners in this field in order to share knowledge and seek assistance
  - Seminar series - again, to enable communications between international colleagues, a seminar series could be another way to share results and feedback between the annual conference cycle
  - Code notebooks - following the model of the classification workstream, we would like to provide additional notebooks covering a wider range of applicable methods to use alternative data sources for consumer price statistics
  - Other uses of these prices alternative data - these data can be used **not just for temporal but also for spatial comparison of prices** as well as for other statistics (for example, scanner data for household expenditure). **UNTT would will start collating guidance on this usage and the ICP team could start a cooperation on this for what concerns PPPs aims .**

# UNTT governance - Secretariat responsibilities

- ▶ The Secretariat is responsible for supporting the task team on the maintenance of the handbook, training materials, code/notebooks, as well as of the task team's website. The Secretariat and nominated leads are set up at the UN Regional Hub in Brazil
- ▶ The Secretariat will operate within a defined framework to ensure the content of the handbook, training materials, code/notebooks remain up to date with international literature
- ▶ It is expected that the Secretariat began their duties in Q2 2025. A consultant (Helen Sands) has been recently hired to support the initial set up of the framework alongside the Secretariat
- ▶ The Secretariat leads are Vladimir Goncalves Miranda and Lincoln Teixeira da Silva from IBGE



Many thanks