



UN Task Team on Scanner Data update

Tanya Flower

UN Regional Hub Webinar - Alternative Data Sources for the CPI

1 July 2025

Virtual

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
- ▶ Other Task Teams include mobile phone data, earth observation data etc
- ▶ <https://unstats.un.org/bigdata/index.cshtml>

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](#)

Summary of workstreams



Handbook

Available [here](#)



Classification

Initial set of guidance released [here](#)



Training

Initial course released [here](#)



System architecture

Interim report released [here](#)



Reproducibility/FAIR

Interim guidance released [here](#)

E-handbook - [LINK](#)

Contents:

- [Acknowledgements](#) — The development of this handbook and associated training material is a testament to the hard work of many individuals who collaborated on drafting, reviewing and updating the content.
- [Glossary](#) — This page contains the handbook glossary.
- [Initial considerations](#) — This section covers some initial considerations that NSOs are recommended to review before setting out to use alternative data sources, such as transaction and web scraped data, in the production of consumer price statistics.
 - [Introduction to the new data sources](#) — This page contains a general introduction to each of the alternative data sources used to compile consumer price indices.
 - [Quality assurance](#) — Quality considerations for a new alternative data sources project
 - [IT system requirements](#) — Requirements for IT systems are dependent on the choice of data source, the implementation plan for each new data source and the availability and knowledge of staff with specialist IT skills.
 - [Tips for optimising computation performance](#) — Methods to optimise computational performance for on-premise computation.
- [Data selection and acquisition](#) — This section discusses how to select and acquire these new data sources.
 - [Selection of categories](#) — One of the initial steps in this process is the identification and selection of product categories to consider for the adoption of alternative data sources
 - [Selection of retailers for alternative data sources](#) — This page contains information on how to identify suitable retailers who can supply these alternative data sources
 - [Scanner data](#) — This section addresses the main steps that should be considered for acquisition of scanner data sets for the production of price statistics.
 - [Initial steps in the acquisition of scanner data](#) — Recommendations on how an NSO can approach retailers to acquire scanner data
 - [Data requirements specification for scanner data](#) — Key variables to ask for when acquiring scanner data.
 - [Data sharing agreements](#) — Data sharing agreements can help provide structure to data flows between a supplier and NSO
 - [Alternative approaches to acquiring scanner data](#) — Suggestions of what to do if the NSO is unable to acquire scanner data directly from retailers
 - [Monitoring, validation and plausibility checks for scanner data](#) — Recommended checks to carry out on the incoming flows of scanner data
 - [Web scraping](#) — This section covers the acquisition of web scraped data.
 - [Different approaches to accessing web scraped data](#) — Approaches for setting up a web scraping project
 - [Strategies available for in-house web scraping](#) — Points to consider if in-house web scraping is the chosen solution
 - [Data requirements specification for web scraped data](#) — Key variables to look for when acquiring web scraped data
 - [Monitoring, validation and plausibility checks for web scraped data](#) — Recommended checks to carry out on incoming flows of web scraping data
 - [Common technical problems for in-house web scraping](#) — Overcoming some common technical problems for web scraping
 - [Example file structures for web scraped and scanner data](#) — This page contains example variables and data types for new data sources
 - [Collection of data via APIs](#) — An alternative to automatically scraping data from web pages is to extract data directly from application programming interface (APIs)
 - [Dealing with unexpected data gaps](#) — This page contains information on how to deal with unexpected gaps in the supply of new data sources

- **Preparation of data** — This section summarises the general steps required to prepare these new data sources for use in price index compilation.
 - **Product sampling for price index calculation** — NSOs can either choose to use all available data from their new data source, or to use a sample of this data.
 - **Standardising the data** — Data will need to be standardised before being used to calculate price indices.
 - **Aggregation across time and outlets** — Data are usually disaggregated by various dimensions. This page contains information on how products can be aggregated across time and outlets.
 - **Identifying unique products** — Data are usually disaggregated by various dimensions. This page contains information on how products can be defined from the given article codes.
 - **Treatment of discounts and refunds** — This page contains information on how discounts and refunds should be treated in the data.
- **Classification** — The goal of this section is to provide guidance on the critical step of classifying new data sources and make them ready for price index compilation.
 - **Pre-conditions and deciding on appropriate classification methods** — There are several factors that NSOs typically consider when approaching or implementing classification, such as the scale and complexity of the task, whether this is just the initial phase of classification or recurrent classification in production, or the structure and quality of the data. Evaluating these factors is useful to select the method(s) most appropriate for each use case.
 - **Note on variables applicable for classification** — Several variables obtained during the acquisition process are typically applicable for classification.
 - **Method 0: Manual labelling or validation of predicted labels** — Manual classification involves each unique product being manually assigned a classification class. This is often described as labelling, and those who specialize in the task, labellers. This method can also be used to validate classification done automatically by other methods.
 - **Method 1: Attribute-based classification method** — Attribute-based classification involves classifying observations in the data by mapping structured variables or attributes available in the alternative data file to the necessary taxonomy classes.
 - **Method 2: Pattern matching classification method** — Pattern based classification involves mapping individual records to the taxonomy classes of interest according to patterns present in their descriptive attributes. For instance, patterns in the categories of the alternative data, patterns in the product name, or specific keywords in the product description - could all be used to classify a set of individual records.
 - **Method 3: Recommendation / Machine-assisted classification** — Machine-assisted classification is still a manual classification technique and therefore the ideas in the manual classification chapter are still valid. The key difference is the addition of a recommendation system that supports labellers in identifying the most appropriate class. Every product is still manually scrutinised, but machine-assistance may improve efficiency and potentially quality by offering recommendations.
 - **Method 4: Machine Learning classification method** — Machine Learning (ML) can be a stable solution for classifying unique products when an NSO starts to use new data sources at scale. NSOs typically use supervised ML methods whereby a manually annotated dataset is used to train a classifier to predict all new unique products as part of recurrent classification.
 - **How to evaluate classification methods** — There are numerous metrics that can be tracked and applied when evaluating how well a classifier performs.
 - **Working with class imbalance** — Datasets NSOs need to classify are typically not equally distributed, as some classes have many records (i.e. unique products) while others have only a few. This requires additional steps to be taken when designing and testing classifiers.
 - **Operational best practices** — Once NSOs select and develop an appropriate classification method, they have to implement it. This section includes articles on appropriate ways to operationalize classification.
 - **Designing the classification step: operational considerations** — This section provides context about where and how the classification step fits into the rest of the pipeline processing alternative data to create elementary price indices, as well as some important considerations and options when setting up the classification process. This article can also be seen as a companion to the pre-conditions to classification section, helping the NSO design the operational process for the individual or blended method that will be chosen.

- [Data filtering and missing prices](#) — This section covers some common methods used to filter the data before it is used to calculate indices, as well as a summary of how missing prices should be treated.
 - [Outlier filter](#) — Outlier filters aim at excluding or correcting extreme price increases or price decreases, typically compared to the previous period, from the price index calculation
 - [Dumping filter](#) — Dumping filters aim at eliminating the downward pressure of clearance prices on the index.
 - [Low sales filter](#) — The low sales filter ensures that products with small expenditure shares do not unduly influence the index results.
 - [Current practice of data filtering](#) — This page contains a summary of how NSOs currently approach data filtering.
 - [Treatment of missing prices](#) — This page contains information on how to treat missing prices through product-level imputation methods.
- [Price index methods](#) — This section summarises the current literature available on price index methods that can be used for these new data sources.
 - [Bilateral price index methods](#) — This section presents the basic, commonly known formulas of bilateral price indices.
 - [Example of chain drift](#) — Chain drift occurs when an index does not return to unity when prices in the current period return to their levels in the base period.
 - [Multilateral price index methods](#) — This section presents the basic, commonly known formulas of multilateral price indices.
 - [Extension methods](#) — Extension methods can be used in combination with multilateral methods to ensure revision-free index series.
 - [Decomposition](#) — In addition to measuring the aggregate price or quantity change, it is important to decompose the change into its contributions by individual commodities or groupings.
 - [Choosing an index method](#) — This section contains information on how an NSO can choose a suitable index method.
 - [Methods for when expenditure data are available](#) — This page summarizes potential criteria that could be used by an NSO to decide on an index method for when expenditure data are available.
 - [Methods for when expenditure data are not available](#) — This page summarises potential criteria that could be used by an NSO to decide on an index method for when expenditure data are not available.
 - [Deriving proxy weights for web scraped data](#) — Web scraped data does not contain expenditure information. In some cases, proxy expenditure weights can be derived.
- [Aggregation](#) — This page contains information on how alternative data sources can be aggregated with existing data sources
- [Other considerations](#) — This section contains information on how to treat seasonal products and quality change.
 - [Seasonal products](#) — This page contains information on the treatment of seasonal products in the CPI and implications of using the new data sources
 - [Quality change and hedonic estimation](#) — This page contains information on accounting for quality change in the measurement of consumer prices
- [Implementation](#) — This page contains information on how to implement these new data sources into production.
- [Training material](#) — This page contains a summary of the planned training curriculum and links to available courses.

E-handbook/ classification

- ▶ The Task Team is currently doing a stocktake of existing handbook content, with the aim of starting to update some of the sections to include more recent findings/ guidance (for example, around the use of bulk web scraping)
- ▶ If you have any suggestions on improving the content - please let me know! tanya.flower@ons.gov.uk
- ▶ Some supplementary guidance will also be available from the classification workstream, along with some example code notebooks to showcase the 4 main methods

Training

- ▶ Our training workstream has designed a curriculum to sit alongside the handbook material to provide additional guidance
- ▶ The first course "Overview of Alternative Data Sources" has been loaded on the UN Learning Platform
- ▶ We've had some technical issues in producing the training material so there has been a delay in uploading the second course on data acquisition to the UN Learning Platform, but we are aiming to have it loaded by mid-July.
- ▶ The next courses will be on index methods and data preparation (aim for Autumn 2025)
- ▶ To access these courses...

Sign up to the [UN Learning Platform](#)...

UN Global Platform - Learning Management System

☐ Remember username

[Forgotten your username or password?](#)
Cookies must be enabled in your browser
Some courses may allow guest access

Log in using your account on:
 UN Global Platform Authentication

Is this your first time here?

For full access to this site, you first need to create an account.

The courses will be listed under
"Alternative Data Sources for CPI"

▼ Big Data

- [What is Big Data?](#) ⁽²⁾
- [Automatic Identification System \(AIS\)](#) ⁽⁴⁾
- [Alternative Data Sources for CPI](#) ⁽¹⁾
- [Mobile phone data](#) ⁽¹⁾
- [Privacy preserving techniques](#) ⁽³⁾
- [Energy Statistics](#) ⁽⁵⁾
- [System of Environmental Economic Accounting \(SEEA\)](#) ⁽¹⁷⁾
- [Social and Demographics Statistics](#) ⁽⁴⁾
- [Administrative Data](#) ⁽¹⁾
- [Sustainable Development Goals \(SDGs\)](#) ⁽⁶⁾
- [Other courses](#) ⁽⁷⁾

Click on "Enrol me"

Overview of Alternative Data Sources

[Home](#) / [Courses](#) / [Big Data](#) / [Alternative Data Sources for CPI](#) / [Overview of ADS](#) / [Enrol me in this course](#) / [Enrolment options](#)

Enrolment options

[Overview of Alternative Data Sources](#)



This introductory course is the first of a curriculum about "Alternative Data Sources for compiling Consumer Price Indices". It aims to raise awareness of what these data sources are and to showcase how they can be applied at a National Statistical Organisation.

The main characteristics of Alternative Data Sources, with their advantages but also challenges for incorporating their use at the National Statistical Institutes, are described focussing the attention on scanner data, web-scraped data, data obtained through Application Programming Interfaces, or APIs, and administrative data.

▼ [Self enrolment \(Student\)](#)

No enrolment key required.

[Enrol me](#)

Watch the videos, complete the Quiz sections and fill in a feedback survey to receive the course certificate...


Overview

Welcome to the "Overview of Alternative Data Sources" course! This introductory course aims to raise awareness of what these data sources are and to showcase how they can be applied at a National Statistical Organisation. It is broken down by six modules:

- Module 1.0: Introduction to Alternative Data Sources, where alternative data sources are introduced.
- Module 1.1: Scanner data characteristics
- Module 1.2: Web-scraped data characteristics
- Module 1.3: Application Programming Interfaces data characteristics
- Module 1.4: Administrative data characteristics
- Module 1.5: Comparison among different ADS and features of an ADS project, that compares the different Alternative Data Sources, described in the previous modules, and sketches the main features of a project for their use for CPI compilation.

This e-learning course and its modules are self-paced and have been developed by the Task Team on Scanner Data of the UN Committee of Experts on Big Data and Data Science for Official Statistics.

In order to receive a course certificate, you must complete the Quiz section at the end of each module with a passing grade, as well as a final [feedback survey](#).

Your progress 

System architecture

- ▶ Thank you to all of you who responded to our survey! We had responses from 70 NSOs from every major geographic region, covering a range of experience with using these new alternative data sources
- ▶ Interim report is available here: <https://un-task-team-for-scanner-data.github.io/production-systems-survey/report-site/>
- ▶ We also presented a summary at the UNECE meeting in Geneva. Recording available here: <https://webtv.un.org/en/asset/k14/k144nqzwye>, it starts at around 09:30 on the recording.

High-level summary of practical suggestions

- ▶ Think explicitly about system boundaries.
- ▶ Think explicitly about data interchange between systems.
- ▶ Embed technical expertise in business domain teams.
- ▶ Adopt Git and GitLab or GitHub as a Version Control System for code, configuration, and documentation.
- ▶ Leverage analytics optimized file formats like Apache Parquet
- ▶ Where it's feasible, implement Complex Analytical Systems as pipelines with one-way data flows and idempotent operations.
- ▶ Practice updating systems more frequently.
- ▶ Ensure your CPI Production System can operate in a separate development environment.

Reproducibility/FAIR

- ▶ 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! Published [here](#)*

We'd love to hear from you!

We're always on the look out for feedback on Task Team work, or even better - for new members to join us!

Please contact the chair or workstream leads directly:

- ▶ Tanya Flower (Chair) tanya.flower@ons.gov.uk
- ▶ Serge Goussev (Classification and FAIR lead) serge.goussev@statcan.gc.ca
- ▶ Federico Polidoro (Training lead) fpolidoro@worldbank.org
- ▶ Collin Brown (System Architecture lead) collin.brown@statcan.gc.ca

What does membership of the UN Task Team entail?

- ▶ You will be allocated to one of the workstreams (partially based on time zone but can take into account preferences as well!)
- ▶ Workstreams tend to meet monthly or as needed to discuss progress and tasks
- ▶ It depends on individual circumstances but the time commitment expected is around 1 day per month

Thanks for listening!

