UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity Dignified Life

New Data Sources & Alternative Price Data Collection Methods

Best Practices in the application of new data sources in Price Collection

Majed Skaini
Amman, Jordan
December 6-7, 2023

01

Introduction

02

Limitations of Traditional Methods

03

Advantages of Modern Methods

04

Web Scraping & Scanner Data

05

Best Practices of New Data Sources

06

Case Studies

07

Limitations & Solutions

08

Future Directions

09

Conclusion

Introduction

Making better use of data is integral to our future and service. We must accelerate a shift in our data and analytics abilities.

António Guterres, UN Secretary-General

Introduction

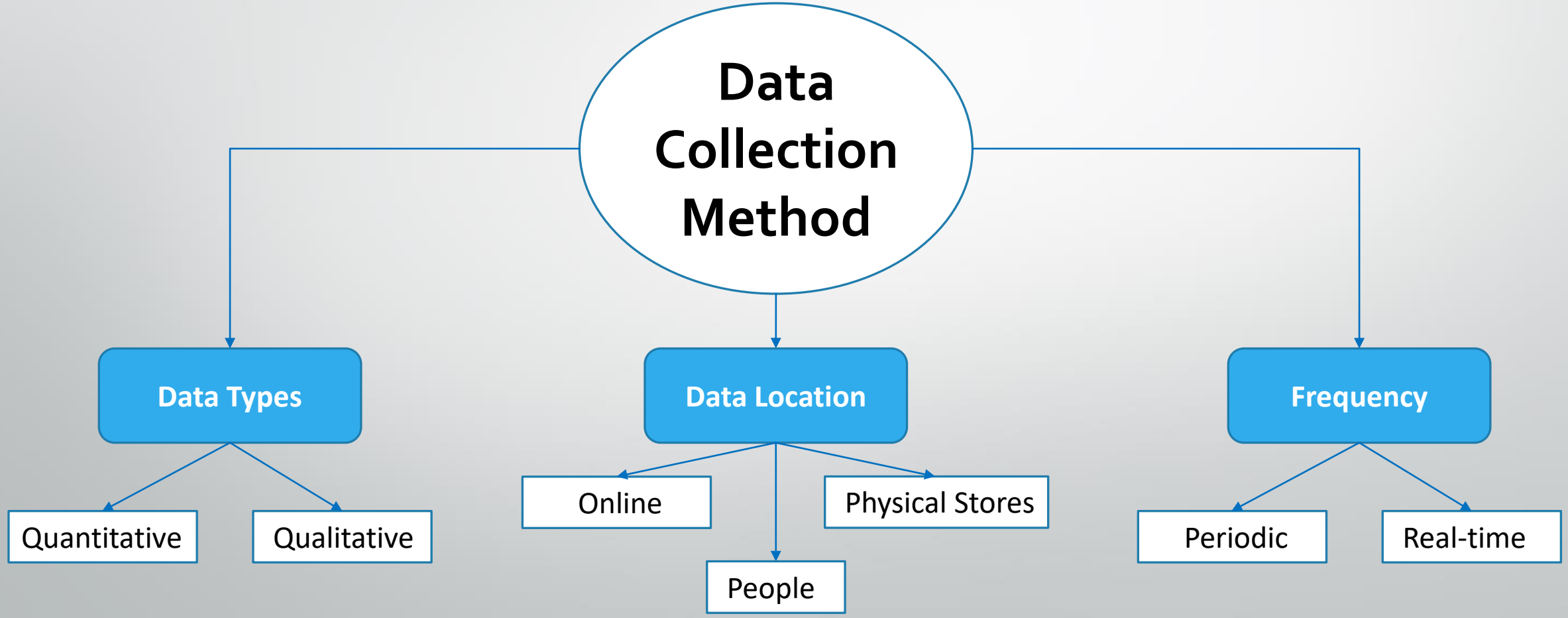
Introduction

The economy is evolving rapidly, influenced by digital transformation, globalization, and changing consumer behaviors. Real-time insights allow national price statistical agencies to capture the dynamic nature of market conditions.

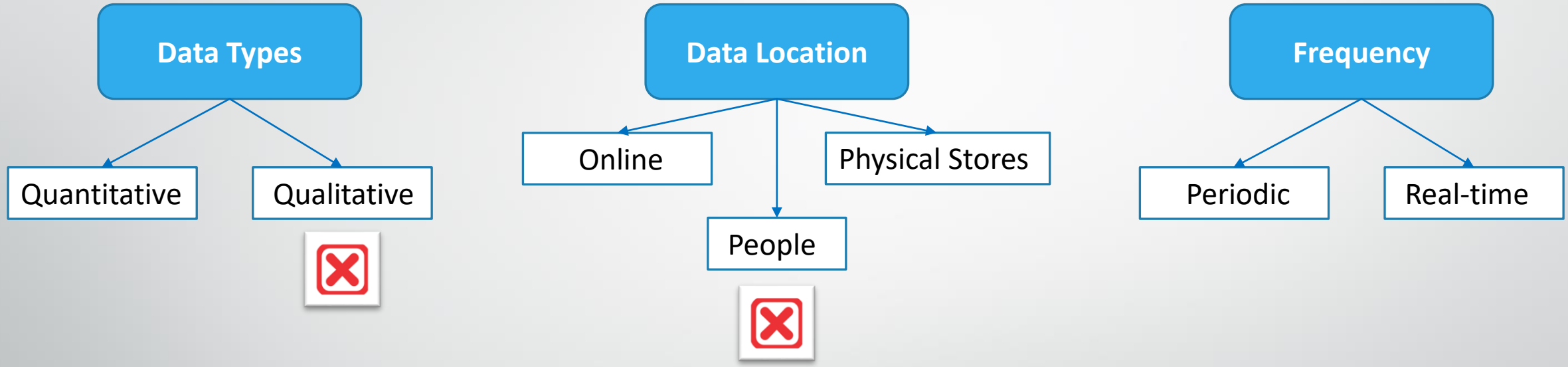
Price indicators are crucial for formulating effective policies that address inflation, economic stability, and overall financial health. These indicators provide valuable insights into the dynamics of prices for goods and services

Data Collection Strategies

There is no superior method for data collection; the choice depends on various factors



Data Collection For NSO



In National Statistical Office (NSO) price data collection, the information is predominantly **Quantitative** in nature, sourced from **online retail platforms** and physical retail **Point-of-Sale** (POS) systems.

The frequency of data collection is flexible, encompassing **both periodic assessments and real-time monitoring**.

02

Limitations of Traditional Methods

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Traditional Methods

Manual Online Data entry

Human operators manually visit websites and record prices.

Manual Price Tag Audits

Personnel physically visit stores equipped with a checklist to manually record item prices

Limitations of Traditional Methods

Inefficiency

Manual data collection methods can be resource-intensive and time-consuming.

↓

Delays in obtaining and processing data, leading to a lag in price reporting and reduced timeliness in reflecting current market conditions.

Limited Frequency & Coverage

Periodic data collection misses real-time fluctuations. Traditional methods also struggle to capture digital transactions and online commerce.

↓

Underrepresentation and misinterpretation of real-time inflation trends.

03

Advantages of Modern Price Collection Methods

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

New Data Sources for Price Data Collection

01

Real-Time Accuracy

Provides an instantaneous and accurate reflection of price changes

02

Efficiency and Timeliness

Automation and advanced technologies significantly reduce gathering of data

03

Enhanced Data Quality

Advanced data processing techniques improve data quality, reducing errors and enhancing the reliability of price data measurements.

04

Web Scraping & Scanner Data

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Scanner Data

Scanner data is revolutionizing the way we collect price data from physical stores. This method involves using point-of-sale (POS) systems, which capture data at the moment of purchase, providing a real-time, granular view of consumer behavior and price changes.

Scanner Data Process

POS Transaction

Customers come to the Point-of-Sale (POS) and scan items for purchase.

Data Storage

Transaction data is saved in the database.

Meanwhile

Legal Framework

Implementation of a legal framework for data sharing between Retailers and NSO.

Data Transfer

The scanner data is transferred periodically depending on the agreed-upon schedule

Importance of Legal Framework

Setting a legal framework between NSOs and Retailers is the most important step in the scanner data process.

1. Data Privacy and Compliance:

Defines how consumer data is handled, ensuring compliance with privacy laws and outlines terms for data collection, sharing, and usage.

2. Clear Terms and Conditions:

Provides explicit terms for data sharing and outlines the rights, scope of data sharing, and responsibilities for data security.

3. Ethical Data Practices:

Ensuring Ethical Conduct: Sets guidelines to prevent misuse or unauthorized access to sensitive consumer information.
Building Trust: Fosters trust between the NSO, retailers, and the public whose data is utilized, crucial for ongoing cooperation.

4. Data Security and Confidentiality:

Specifies security measures, including encryption and access controls, to protect data confidentiality and integrity.

Legal Framework between NSOs and Retailers

- Sample -

CONFIDENTIALITY AGREEMENT

Agreement number: 2013/ H

BETWEEN

THE BELGIAN STATE (THEMATIC DEPARTMENT PRICES OF STATISTICS BELGIUM OF THE FPS ECONOMY, SMES, SELF-EMPLOYED AND ENERGY)

AND

THE SUPERMARKET CHAIN

The Belgian State (Thematic Department Prices of Statistics Belgium of the FPS Economy, SMEs, Self-Employed and Energy), with registered office in *xxx*, Enterprise number: 0314.595.348, hereinafter called 'Statistics Belgium', for the purpose of this agreement represented by xxx, Deputy Prime Minister and Minister of Economy.

And

The supermarket chain, address, for the purpose of this agreement represented by *XY*

Observe that:

- In order to calculate the consumer price index, as well as to meet its European obligations with regard to the harmonised index of consumer prices, the project average prices and purchasing power parities, Statistics Belgium is required to carry out price observations in shops that are part of *the supermarket chain*;
- Statistics Belgium tries to use as many electronic files on sales and turnover at product level as possible for the statistics mentioned above,
- in order to further improve the reliability of these statistics;

Have agreed:

Article 1. Transmission of data to Statistics Belgium

1.1. *The supermarket chain* shall transmit the data only to Statistics Belgium.

1.2. *The supermarket chain* shall transmit data to Statistics Belgium in electronic form each week for the production of the statistics mentioned above. The transmitted data shall at least be broken down by week, i.e. from Monday to Sunday.

1.3. To ensure the timely production of these statistics, the data mentioned in article 2 shall be transmitted at the latest on the second working day of the week following the week to which the data refer. A working day is understood to be every day of the week, excluding Sundays and bank holidays.

1.4. Data shall retroactively be transmitted from 1 January 2012, these historical data shall at least be broken down by month. Data shall be transmitted through a secured connection.

Article 2. Content of the transmitted data

2.1. The data to be transmitted to Statistics Belgium are defined in annex 1 of this agreement.

2.2. If *the supermarket chain* makes a selection of the sold products in the data, these will be provided for all products from its range that can be traced to the product groups in annex 2.

2.3. The data for each product may be aggregated for all points of sale of *the supermarket chain* combined. All points of sale are understood to be those for which *the supermarket chain* can provide data.

2.4. If *the supermarket chain* comprises multiple retail formulas, the data shall be broken down by retail formula wherever possible. For the application of this agreement, *the supermarket chain* shall provide data for the following retailers: ….

2.5. Reporting comprises both goods that are bought directly at the shop as well as distance selling.

2.6. The provided data only refer to sales to consumers.

Article 3. Use of the data by Statistics Belgium

3.1. Statistics Belgium shall use the data exclusively and only (1) for the calculation of the consumer price index, as well as to meet its European obligations with regard to the harmonised index of consumer prices, the project average prices and purchasing power parities and (2) to deliver a market report to Comeos in which aggregated volume and price data for product groups as listed in annex 2 are provided.

3.2. Based on the data, Statistics Belgium shall produce aggregated statistics to fulfil its task of providing official information for public use. To succeed in this task as well as in related tasks, it shall meet the provisions included in the law of 4 July 1962 on public statistics.

3.3. In accordance with the law of 4 July 1962 on public statistics, Statistics Belgium shall ensure that no information related to the individual situation of *the supermarket chain* can be derived from the data published by the Thematic Department Prices of Statistics Belgium.

3.4. If the delivered data fall within the scope of article 111 of the NAI law containing social and various provisions (21 December 1994), the agreement shall be legally void.

Aricle

Article 4. Transmission of the data to third parties

4.1. The supermarket chain shall transmit the individual data referred to in article 1 exclusively to Statistics Belgium.

4.2. Statistics Belgium may not transmit the individual data to a third party.

4.3. The supermarket chain shall give its permission to use the provided data for an aggregated report made available to Comeos (not-for-profit organisation).

4.4. The report may not be disseminated as an official statistic, since it is not representative of the entire market.

4.5. Statistics Belgium shall carry no responsibility for the use of the report by Comeos. In case of disputes related to the use of this report, the supermarket chain shall turn to Comeos.

Article 5. Unforeseen circumstances

5.1. If - due to unforeseen circumstances - the electronic data cannot be transmitted in time, *the supermarket chain* shall inform Statistics Belgium of the reason and expected duration of the delay one working day after the period referred to in article 1.2.

Article 6. Revision

6.1. Each party to this agreement may request its revision.

6.2. All changes to this agreement shall occur by means of a written annex, which will be agreed on under the same conditions as this agreement. A verbal agreement between parties is not binding.

Article 7. Duration and termination of this agreement

7.1. This agreement shall go into effect on *1 September 2013* and shall be subject to a trial period of 3 months. After this trial period the agreement will automatically be renewed each year, unless one of both parties in writing asks for the termination of the agreement, at least 3 months before the end of the duration of the ongoing year.

7.2. The first year referred to in article 7.1 runs from the end of the trial period to 31 December 2014.

7.3. Each shortcoming by Statistics Belgium with regard to the obligations in this agreement grants *the supermarket chain* the right to end the agreement with immediate effect, after Statistics Belgium has been delivered a notice of default by registered letter to remedy the shortcomings, and if this notice of default has not yielded an adequate outcome at the latest 8 working days after its sending.

Article 8. Applicable law and competent courts

8.1. This agreement shall be governed only by Belgian law.

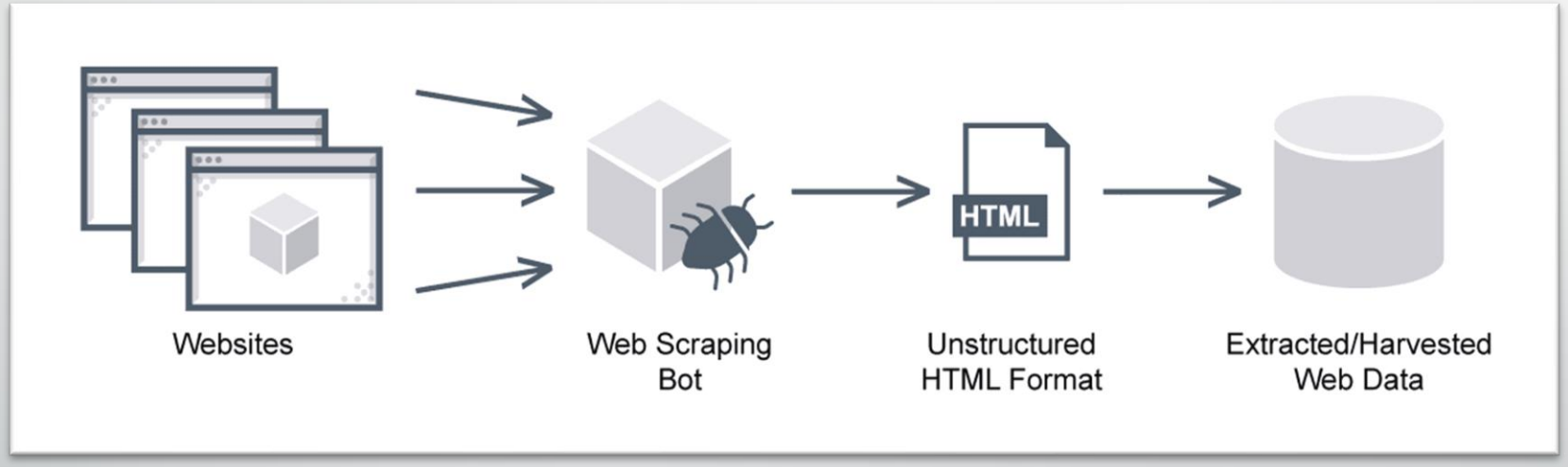
8.2. If, due to circumstances or occurrences, problems arise with the execution or interpretation of this agreement, the parties shall commit to trying to reach an agreement that meets the demands of both parties and that is in accordance with the law of 4 July 1962 on public statistics and its implementing decisions, before any other steps are taken.

8.3. Any dispute arising from this agreement shall be subject to the exclusive jurisdiction of the courts of Brussels.

Done in Brussels on (date) in as many original copies as there are parties to the agreement. Each party acknowledges to have received an original copy.

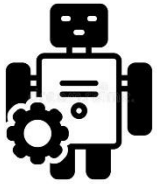
Web Scraping

It is the process of extracting data from websites. It has gained prominence as an innovative solution for improving price data collection. With web scraping, it becomes possible to access and collect real-time pricing information from a vast array of online retailers.



This technique offers the ability to monitor price fluctuations in the digital space, enabling a more accurate representation of changing prices

Benefits of Web Scraping



Automation

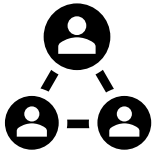
Automatic extraction for websites data rather than manual pull out

⇨ *Time saving*

Accuracy

Data is accurately extracted from websites into spreadsheets

⇨ *Eliminating any human error*



Efficiency

Less resource-intensive than traditional prices collectors

⇨ *Cost effective*

Modern Price Data Collection Methods

Web Scraping

Scanner Data

Both web scraping and scanner data are illustrative of how modern data sources are adapting to the needs of an evolving economy, promising more precise and dynamic price indicators measurements that covers both the Digital & Physical spaces.

05

Best Practices of New Data Sources

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Best Practices of New Data Sources

Best Practices	Web Scraping	Scanner Data
Data Quality and Accuracy	Design precise scraping algorithms for specific data extraction.	Regularly validate and verify that scanner data accurately reflects point-of-sale transactions.
Data Processing and Cleaning	Standardize product names, categories, and units of measurement during data processing.	Identify and rectify missing or duplicate entries. Clean data for consistency.
Privacy and Ethical Considerations	Check a website's crawling permissions to respect ethical scraping practices.	Anonymize or aggregate data to protect individuals' privacy.

Best Practices of New Data Sources

Example

Privacy & Ethical Considerations

'robots.txt' is a standard used by websites to communicate with web crawlers and other automated agents about which parts of the site should not be crawled or processed.

To access the file, simply add "/robots.txt" to end of any website. Example:

https://www.example.com/robots.txt

```
User-agent: *
Allow: /researchtools/ose/$
Allow: /researchtools/ose/dotbot$
Allow: /researchtools/ose/links$
Allow: /researchtools/ose/just-discovered$
Allow: /researchtools/ose/pages$
Allow: /researchtools/ose/domains$
Allow: /researchtools/ose/anchors$
Allow: /products/
Allow: /local/
Allow: /learn/
Allow: /researchtools/ose/
Allow: /researchtools/ose/dotbot$

Disallow: /followerwonk/bio*
Disallow: /products/content/
Disallow: /local/enterprise/confirm
Disallow: /researchtools/ose/
Disallow: /page-strength/*
Disallow: /followerwonk/profiler/*
Disallow: /thumbs/*
Disallow: /api/user?*
Disallow: /checkout/freetrial/*
Disallow: /local/search/
Disallow: /local/details/
Disallow: /messages/
Disallow: /content/audit/*
Disallow: /content/search/*
Disallow: /marketplace/
Sitemap: https://moz.com/blog-sitemap.xml
```

06

Case Study

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Case Study

United States - Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) has stated its strategic objective to "convert a significant proportion of the CPI [Consumer Price Index] market basket from traditional collection to nontraditional sources and collection modes, including harnessing large-scale data, by 2024" (BLS presentation to the panel, October 7, 2020).

level at which prices are aggregated. Current methods are briefly reviewed, then alternative data sources—focusing on various types of scanner and web-scraped data—are assessed for their potential to improve the accuracy, coverage, and timeliness of elementary indexes. Challenges to

stores.[12] By implementing computer-assisted data collection methods over the years, BLS has done an admirable job mitigating timing and accuracy problems with estimating price relatives for the elementary indexes.

Case Study

Australia - The Australian Bureau of Statistics (ABS)

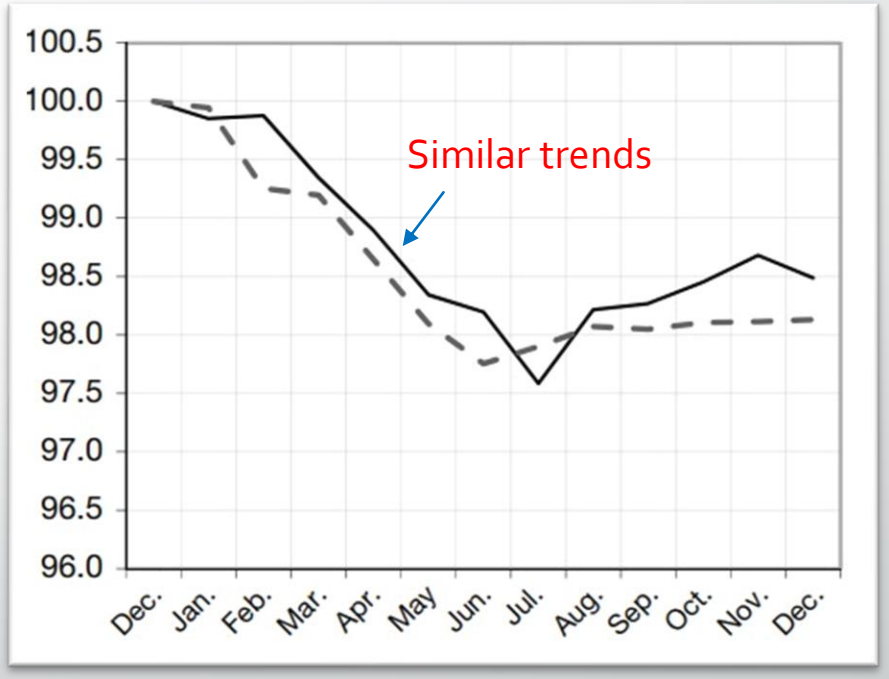
ABS uses scanner data from retailers to obtain prices for about 16 percent of Australia's CPI by item weight. Covering approximately 84 percent of all expenditures at supermarkets,

ABS has been incorporating web-scraped prices progressively into its CPI since March 2017, currently using primarily a direct replacement strategy. It is the primary data approach for some significant item categories such as alcohol and tobacco (7.3 percent weight), clothing and footwear (3.5 percent), furniture and household equipment (3.7 percent), and recreation and

Case Study

France

Since 13 April 2017, an order signed by the minister for the economy requires non-specialised retailers with space allocated to food and drink products of at least 400m², to share in-store scanner data. This facilitates and ensures access to scanner data, which is a prerequisite for compiling an index such as the CPI, which is produced within short time frames and cannot be revised.



The alignment suggests that the Scanner Data Index is a reliable and representative.

Consumer price indices for item headings and indices calculated using scanner data, 2014

06

Limitations & Solutions

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Scanner Data Limitations & Solutions

1. Greatly dependent on the retailers to provide the data, and retailers might be reluctant to provide scanner data to NSOs
   - ✓ Establishing a legal framework
   - ✓ Securely automating data transmission

2. Processing huge amount of data that needs certain validation and IT infrastructure and staff
   - ✓ Having adequate IT infrastructure and specialized/trained staff on both sides
   - ✓ Validating data using the quality framework

Web Scraping Limitations & Solutions

1. Web designs are always evolving, making it harder to label data for scraping
   - **Customize and update the script to handle complex web designs**
2. Certain websites may prohibit data extraction or use of bots fearing increase in traffic
   - **Contact websites and inform them of the actual intentions to avoid legal issues**

07

Future Directions

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Future Directions

Artificial Intelligence (AI) is set to completely transform how we collect, process smarter, faster, and more adaptable.

Data Acquisition

Using Natural Language Processing (NLP) techniques into the scrapers; permitting the codes to scrape data only for the desired items. This approach hugely reduced the scraping execution time

Faster Data Processing

AI speeds up the process of sorting and understanding data, giving us insights in real-time and helping us react quickly to changes.

Adapting to Website Changes

Automatically adjust to changes in websites' HTML file, making it more flexible and less reliant on manual adjustments

09

Conclusion

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

Conclusion

Modern price data collection methods, such as web scraping and scanner data, emerge as a catalyst for redefining how we perceive and measure economic phenomena. The integration of these methods into price statistics significantly enhances accuracy and timeliness. Beyond mere improvements, this evolution paves the way for a future where more advanced data collection methods seamlessly integrate into economic measurement

THANK YOU

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**