VOORBURG GROUP
OTTAWA CANADA
13 | 15 | 20 | 22
SEPTEMBER
2022

Guidelines for Incorporating Alternative Data Sources in Official Statistics

37th Voorburg Group Meeting
September 13-22, 2022

Virtual Meeting hosted by Canada

Anthony Dawson

An Phríomh-Oifig Staidrimh / Central Statistics Office

Rohan Draper

DANMARKS STATISTIK

Scott Kilbey

Office for National Statistics

Martin Beaulieu
Kyle Virgin

Statistics Canada / Statistique Canada

VOORBURG GROUP
OTTAWA CANADA
13 | 15 | 20 | 22
SEPTEMBER
2022

Guidelines for Incorporating Alternative Data Sources in Official Statistics

This Voorburg Group Task Force was formed to:

> *"Build on the results of the 2021 poster session on criteria for fitness of use of alternative data to establish a global guideline for alternative data use"*

Approach:

- Determined the challenges faced by NSIs in assessing the quality of new data sources
- Aligned alternative data literature and practical experiences with the **Generic Statistical Business Process Model (GSBPM)** to build consistency in quality evaluation
- Provided **practical tool** to help National Statistics Institutions navigate the data journey and address risks to data quality along with measures and mitigations

OTTAWA CANADA
13 | 15 | 20 | 22
VOORBURG GROUP SEPTEMBER 2022

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Key Literature referenced:

GSBPM - Generic Statistical Business Process Manual (version 5.1) - https://statswiki.unece.org/display/GSBPM/Clickable+GSBPM+v5.1

Guide to reporting on administrative data quality (Stats NZ, 2020) - https://www.stats.govt.nz/methods/guide-to-reporting-on-administrative-data-quality

Total Survey Error: Past, Present and Future (Groves, Lyberg, 2010) - https://academic.oup.com/poq/article/74/5/849/1817502

Voorburg Task Force - Alternative Data Sources (Aizcorbe et al. - Voorburg, 2020) - https://www.voorburggroup.org/Documents/2020%20Helsinki/Papers/2018.pdf

Voorburg Group 2021 Poster Session (Hutchinson et al. - Thursday 23rd Sept) - https://www.bls.gov/voorburg-dc-2021/agenda-documents/

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Key considerations:

- Alternative Data ≠ Administrative Data
- External vs Internal Data
- Structured vs Unstructured Data
- High Control vs Low Control
- Data Ethics

All of the above consideration should influence how an NSI incorporates a new data source into their statistical processes. It is important to acknowledge these differences and to tailor the practical tool around the data source.

OTTAWA CANADA
13 | 15 | 20 | 22
VOORBURG SEPTEMBER
GROUP 2022

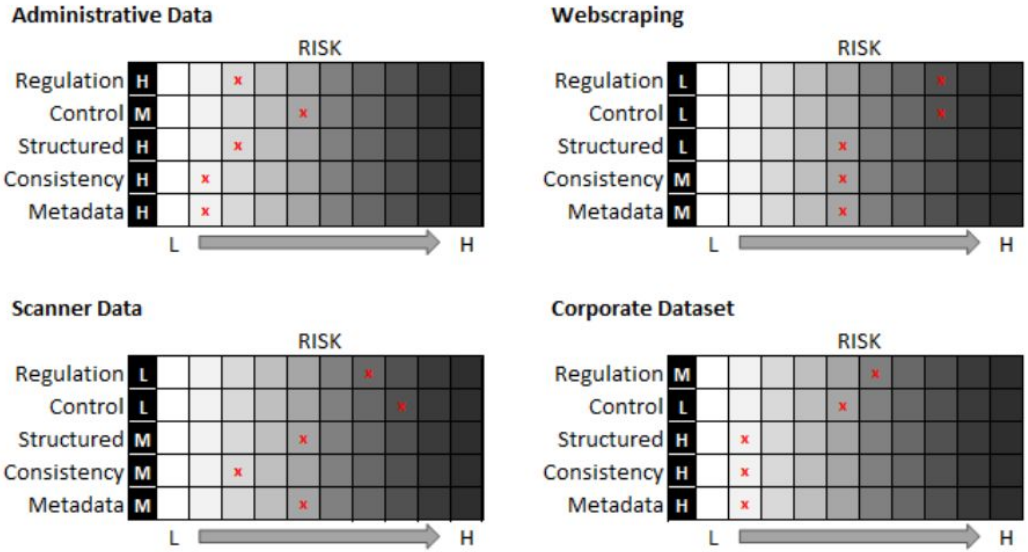
Guidelines for Incorporating Alternative Data Sources in Official Statistics

Source	Type	Origin	Code	Description
TRADITIONAL	STRUCTURED	**Questionnaire** (paper, phone and/or electronic)	QNR	The traditional way of collection price information by asking firms for the information via telephone, paper or electronic questionnaires. This is the default and not an alternative data source.
EXTERNAL	UNSTRUCTURED	**Web Prices** (manual)	WPR	The price collection data is derived from websites manually. The data source itself is considered an unstructured data source not designed for statistical purposes.
		Webscraping (automated)	WSC	The price collection data is derived from websites automatic systems. The data source itself is considered an unstructured data source not designed for statistical purposes.
	STRUCTURED	**Administrative Data Source**	ADM	Data which are derived from the operation of administrative systems by public agencies (e.g. data collected by government agencies for the purposes of registration, transaction, regulation and record keeping). Data is often structured for administrative purposes and is highly transferable for statistical purposes.
		Corporate Datasets	COR	Survey respondent provided datasets obtained directly from corporate headquarters in lieu of data collectors collecting data in respondent stores or on their websites. Data pertains to the particular company that is providing said data is often structured for organisational purposes and is highly transferable for statistical purposes.
		Trade Associations	TAD	Industry based surveys that the target industry is producing for themselves.
		Data Vendors (commercially available structured data)	DVS	Data acquired from companies that actively collect and sell data as a business activity. Often such companies provide data on a contractual basis with defined terms and conditions.
		Consultancies (mandated specific task) (transformed data)	CON	Consulting company and/or specialist company is contracted to collect and/or compile data for a specific purpose (mandated or otherwise). Often such companies are utilised on a contractual basis with defined terms and conditions.
		Credit card and bank data	CCD	Financial information collected at the moment of a transfer of funds between a card holder's account and a business account. Data is graded based on the level of metadata available about the transaction. This source is considered a structured data source.
	BOTH	**Other alternative data sources n.e.c.**	OTH	Other types of alternative data sources not elsewhere classified. For example, transaction-level data from email receipts (like UBER email receipt data). Other special data delivery from third party data collectors not elsewhere classified.
INTERNAL	STRUCTURED	Consumer Price Index	CPI	Data is sourced directly from the Consumer Price Index
		Producer Price Index	PPI	Data is sourced directly from the Producer Price Index
		Structural Business Statistics	SBS	Surveys utilised for benchmarking purposes
		National Accounts	NA	Price indices derived from volume and value data (implicit price indices)

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Variety of Data Sets & Perceived Risks

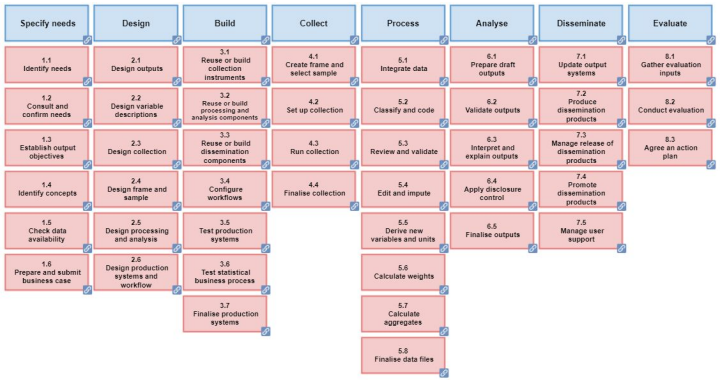
Countries should consider the array of possible data sources and the perceived risks that accompany them



Guidelines for Incorporating Alternative Data Sources in Official Statistics

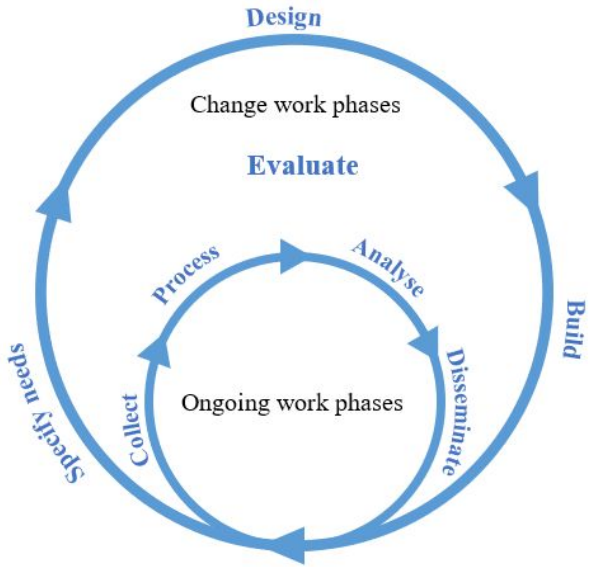
Generic Statistical Business Process Model

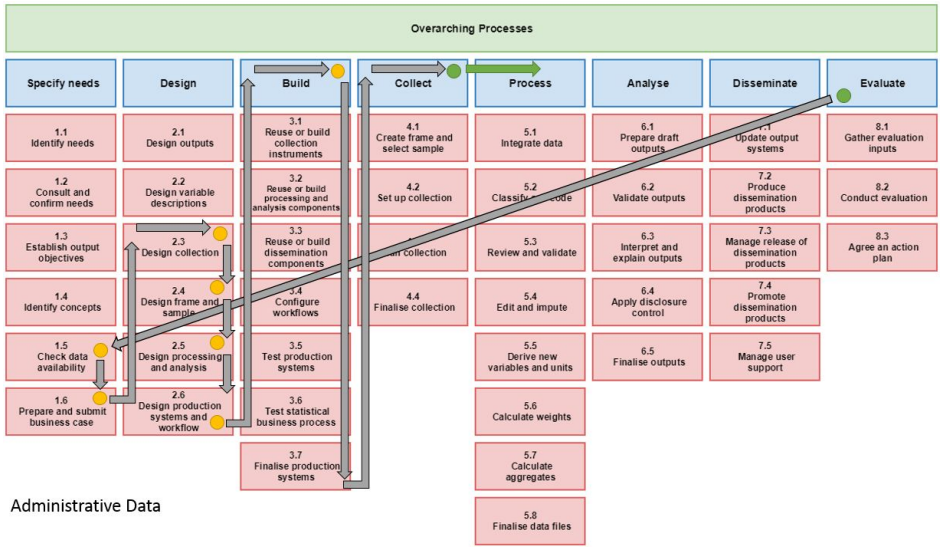
- Version 5.0 of the GSBPM was updated *"to be less survey-centric and [include] activities related to working with non-statistical data providers"*
- *"it is not a rigid framework in which all steps must be followed in a strict order"*
- *"The elements of the model may occur in different orders in different circumstances"*
- *"Some sub-processes will be revisited, forming iterative loops"*



Guidelines for Incorporating Alternative Data Sources in Official Statistics

The integration of alternative data sources takes advantage of these characteristics of the GSBPM

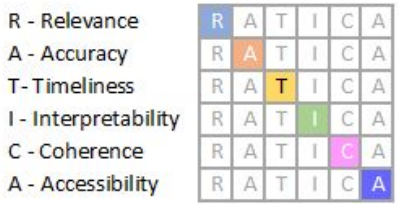




Guidelines for Incorporating Alternative Data Sources in Official Statistics

Fitness for Use Questionnaire

- Questions follow the **GSBPM structure**
- **Suggests measures** that can be taken to assess and mitigate risks to quality and sustainable statistics
- Each question associated with **predominant quality dimension**, if relevant



Guidelines for Incorporating Alternative Data Sources in Official Statistics

Fitness for Use Questionnaire: Example 1

Quality Dimension

4 - Collect

GSBPM

R A T I C A

Question

a. Does the collection process (performed by the data provider) of the admin data file have any impact on the intended use? If so, are there any means that can be used to mitigate or eliminate this impact?

i. Coverage rate of the database used
ii. SNZ metric 7 (Measurement): Percentage of records from proxies
iii. Total and partial response rate (see also Stats NZ metric 5: Item non-response & Stats NZ metric 23: Unit non-response)
iv. Refusal rate
v. Impact of follow-up strategies
vi. Impact of collection mode (suggestion: mode effect)
vii. Capture or coding error rate
viii. SNZ metric 13 (Processing): Percent of transcription errors

Measures & Mitigation

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Fitness for Use Questionnaire: Example 1

5 - Process

e. Are standard concepts and/or classifications being used in the data file? If not, how will this be addressed?

RATICA

i. *A data dictionary and a user guide are available, as needed*
ii. *Detailed description of the main statistical concepts, including statistical measures, population, variables, units, domains and reference period*
iii. *Accurate references for the concepts, variables and standard classifications used*
iv. *Record percentage of items in data set that deviate from target concepts and/or classifications. Note - this may change over time if the data set is unstructured/dynamic*
v. *SNZ metric 1 (Validity): Percentage of items that deviate from target concept definition*
vi. *SNZ metric 2 (Validity): Percentage of items that deviate from international standards or definitions*

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Discussion & Future Work

The task force recommends that:

1. This discussion be added to the agenda for Voorburg 2023
2. Member countries experiment with the proposed questionnaire approach using existing and new alternative data sources in order to provide feedback on it utility

Guidelines for Incorporating Alternative Data Sources in Official Statistics

Feedback from Member Countries

The tool developed by the task force leaned on:

1. GSBPM
2. Statistics Canada Quality Guidelines
3. Guide to reporting on Administrative Data Quality (New Zealand)
4. Internal documentation from Statistics Canada (Quality Evaluation Questionnaire)
5. The experience of Task Force members

We expect there are more valuable resources that can further this work and that this tool will evolve.

This is why we encourage the input of other Voorburg participants to incorporate the variety of data sets, practices and experiences that they bring.