The Importance of Data Quality

Quality and Data Ethics Secretariat
Statistics Canada

May 24, 2022

Delivering insight through data for a better Canada

Outline

- Quality and Data Ethics Secretariat
- 6 dimension of quality
- Fitness for use
- Quality Governance at Statistics Canada

Quality and Data Ethics Secretariat

Our mandate:

- Promote quality processes and produce relevant information from high-quality, ethically sourced data
- Conduct ethical reviews and make recommendations to Chief Data Ethics and Scientific Integrity Officer

Key terms to know for this presentation

Data	Information

• Facts and figures
• Available from many sources
  • Surveying respondents directly
  • Administrative files (e.g. billing files from service providers)
  • Web scraping (Big Data)
• Available in many formats (lists, maps, satellite images, etc.)
	• Making sense of the data
	  • Or, the ability to tell a story about your data through text or data visualization (graphs, infographics)
	• Information about data (or information about information ☺) is called **metadata**

Data should be **ethically sourced**: collected in a transparent manner, and used in a meaningful way that will not cause harm to the respondent

Quality Assurance vs. Quality Control: What's the difference?

Quality Assurance (QA)	Quality Control (QC)
• Should be considered at the **planning** stage • To anticipate problems before they happen • Preventative action • Applicable to all stages of the statistical process / data journey	• Carried out during the **production** stage • To react to observed problems • Corrective and/or preventative action • Applicable to certain stages of the statistical process / data journey

What is quality assurance (QA)?

Everything we do to make sure our work is of good quality

- Meeting the needs of the recipient of our work
  - Everyone from data users to higher management to your teammate ☺
- Effectively managing our time and effort
- Providing clear and thorough documentation
  - Seamless transition between you and your successor/backup
- Not relying on hope, luck, or somebody else to get the job done right
- Proactively mitigating risks to quality

What is quality control (QC)?

- A gate check to measure the quality performance of a process or product
  - Applies everywhere: completeness of a data file, success of a record linkage process, formatting of output file
  - Each time, we compare the measurement to a pre-set standard
  - Make a decision: leave the process alone, or intervene to adjust/correct it
- Traditional examples:
  - Printing paper questionnaires
  - Monitoring screen shots/views
  - Observing phone/in-person interviews
  - Data capture and coding

DATA QUALITY: IT'S A PIECE OF CAKE!

Accessibility

Users are aware of/can access the data

- **Organized**: A system or catalogue is in place to allow users to locate all available data
- **Available**: Once the desired data source's location has been determined, there must exist a channel of communication/distribution plan between the provider and the user
- **Robust**: The catalogue system must be maintained and kept up-to-date
- **Accountable**: Service standards are in place to assist users experiencing difficulty/dissatisfaction with any aspect of data access
- **Affordable**: A balance between freely accessible data and cost recovery for provision of special requests

Accuracy

The data source (and/or estimates) reflects the measure of interest

- Representative of the required population, reference period, etc.
- Methods (collection, methodology) are transparent
- Data are produced without bias or interference
- Reasonable coefficient of variation/confidence interval
- Any under-/overcoverage is known and has been addressed

Estimates refer to statistical inferences about a population measure (e.g. unemployment rate, population size by age group, etc.)

Coherence

Data are consistent over time, between regions and across sub-populations

- An easy way to remember coherence is the three **C**s:
  - **Consistency:** Use standard frameworks, codesets, concepts, variables and classifications
  - **Commonality:** Use common frames, methodologies and systems for data collection
  - **Comparability:** Minimize changes made to standards, classifications and the coding of variables to maintain comparability with the past

Interpretability

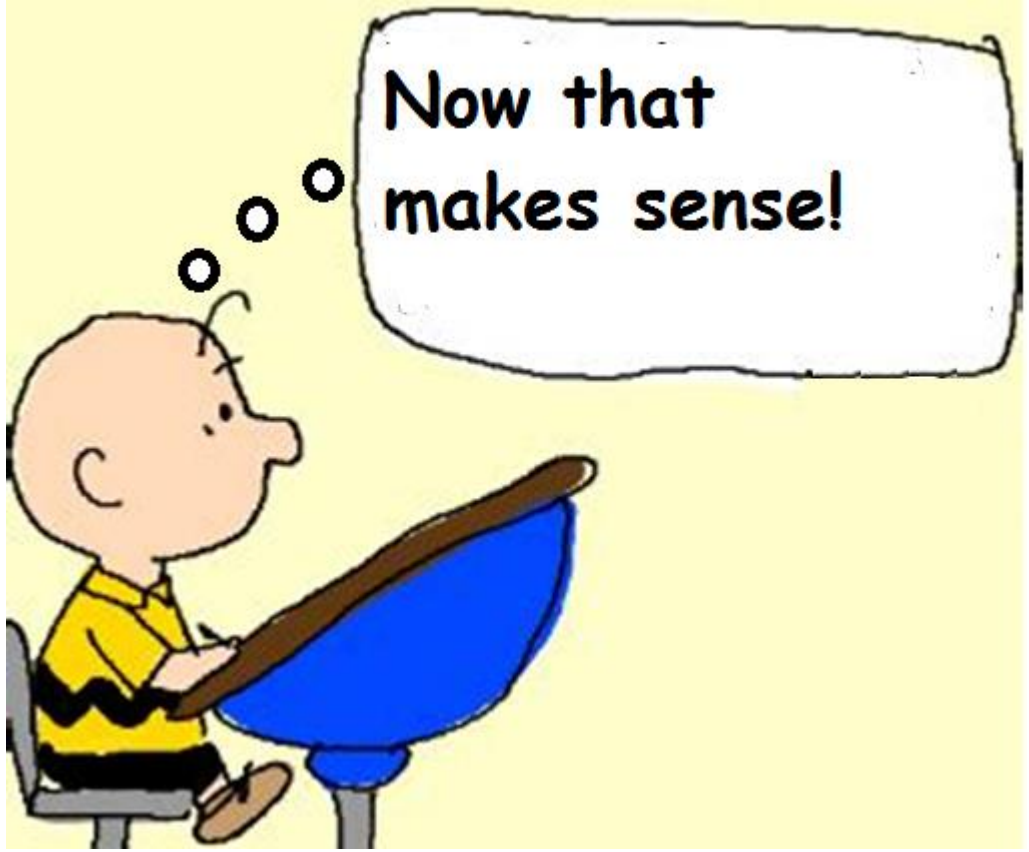
Metadata are available and complete

- Metadata explain the underlying concepts and classifications used to create the data product
- It is used not only to interpret ***how*** best to use a data source, but to determine whether it ***is*** the best data source available for one's specific purposes.

Interpretability (cont'd)

Quality metadata should be…

- Relevant and useful, covering the important elements
- Timely: Available when the data to which it pertains are available
- Complete, accurate, and reliable
- Formatted to recognized standards
- Accessible and understandable



Relevance

Does the information matter to Canadians?

- The relevance of a data product can be tested with four questions:
  1. Is it useful in building policy?
  2. Does it aid in long-term planning?
  3. Does it fill a data gap?
  4. Can it promote new initiatives?

- With long-term planning, we can remain on track/budget while keeping pace with the ever-evolving data needs of users and stakeholders
- However, the scope of what is relevant is constantly changing

Timeliness

The lag between the data source's reference period and its availability for use

- How important is speed to you?
  - Are you willing to accept a lower coverage rate (accuracy) to get the data faster?
- Acceptable gap depends on scale of the source
  - For example, a monthly data cycle should have a shorter grace period than an annual one

Should we bother with quality?: A case study

- A data gap has been identified: Information is needed on the **average dollars spent per month** by Canadians on overnight travel accommodation.
  - Monthly estimates are needed at the province and age group levels
  - Conducting a survey is costly and could be burdensome to travelers

- However, there is good news: You have learned about the administrative data file "The Monthly Register of Canadian Overnight Stays"

What could go wrong?

Back to the case study…

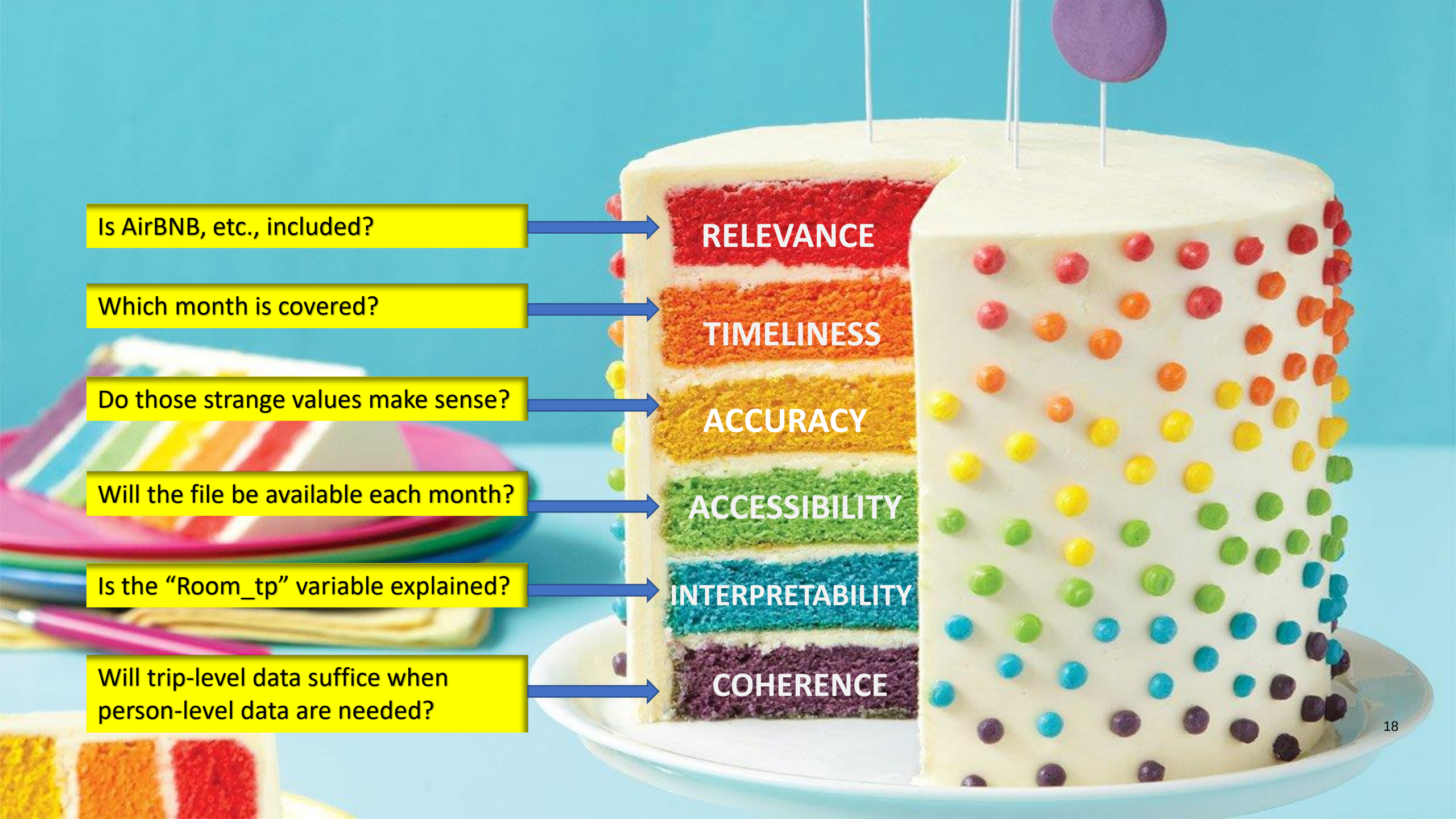
Here is “The Monthly Register of Canadian Overnight Stays” data file*

- Actual file contains 42,985 rows: only certain rows are displayed

Line_no	Establishment	Prov	Room_tp	Nights	CDN_spent
00001	Fisherman's Bed and Breakfast	NL	B	3	30
00250	Cozy Cove Motel	NS	A	2	250
02555	La belle auberge	QC	A	5	600
10800	Capital Inn	ON	A	7	950
24490	Regina Resort	SK	C	1	2500
35590	Prairie Place	AB	D	5	5
42985	Western Hotel	BC	D	7	1200

Thinking back to the six dimensions of quality, **what are some potential issues to consider** before using this file as a survey data replacement?

*All businesses on this file are entirely fictional. In fact, the file itself is fictional.


Is AirBNB, etc., included?
Which month is covered?
Do those strange values make sense?
Will the file be available each month?
Is the “Room_tp” variable explained?
Will trip-level data suffice when person-level data are needed?
RELEVANCE
TIMELINESS
ACCURACY
ACCESSIBILITY
INTERPRETABILITY
COHERENCE

Fitness for Use

- In practice, a data file is unlikely to fail as miserably as “The Monthly Register of Canadian Overnight Stays”
- The data file in question may not be perfect, but that does not automatically mean that it should be abandoned
- The decision of whether to proceed comes down to **fitness for use**

So what exactly is “Fitness for Use”?

There is no single correct way to determine a data source’s overall quality. Rather, the question of assessing quality comes down to the needs of the user. This concept is known as Fitness for Use.

Before exploring available data sources, you should decide:

- What the data source **must have**
- What it would be **nice to have** in the data source

Fitness for Use: A case study

The problem: You need data to produce your city's annual snow removal budget for five years.

Where do you start?

Fitness for Use: Snow removal

Quality dimension addressed	What the data source must have	What would be nice to have in the data source
Relevance	City-level snowfall data	Neighbourhood-level snowfall data
Timeliness	Snowfall data from the previous season	Updated data during the snow season
Accuracy	Values in line with expectations	Measured readings from a trusted source
Accessibility	Available before preparing this season's budget	Available for future years
Interpretability	A description of all variables and an available contact person to provide assistance (robust metadata)	
Coherence	An assurance that file updates will measure the same variables of interest for the same area	

Quality Evaluation

- Choosing between multiple data sources:

Required characteristic	Source 1	Source 2	Source 3
City-level snowfall data	Province-level	City-level	Neighbourhood-level
Available in time for annual budget	Updated annually each December	Updated monthly	Updated annually each August
Accuracy of data	Measured readings from a trusted source	Measured readings from a trusted source	Measured readings from a trusted source
Meets requirements?	✗	✓	✓

What is at risk when quality is ignored?

- **Accuracy**: Without proper quality vetting, data can be inaccurate or misleading
- **Loss in efficiency**: Good quality processes can/should be shared between programs, creating time savings down the line
- **Loss of money**: Getting it right the first time is more cost-effective than corrections or recalls
- **Trust**: Public and internal trust can erode when information is not trustworthy

Don’t let your data become the source of fake news!

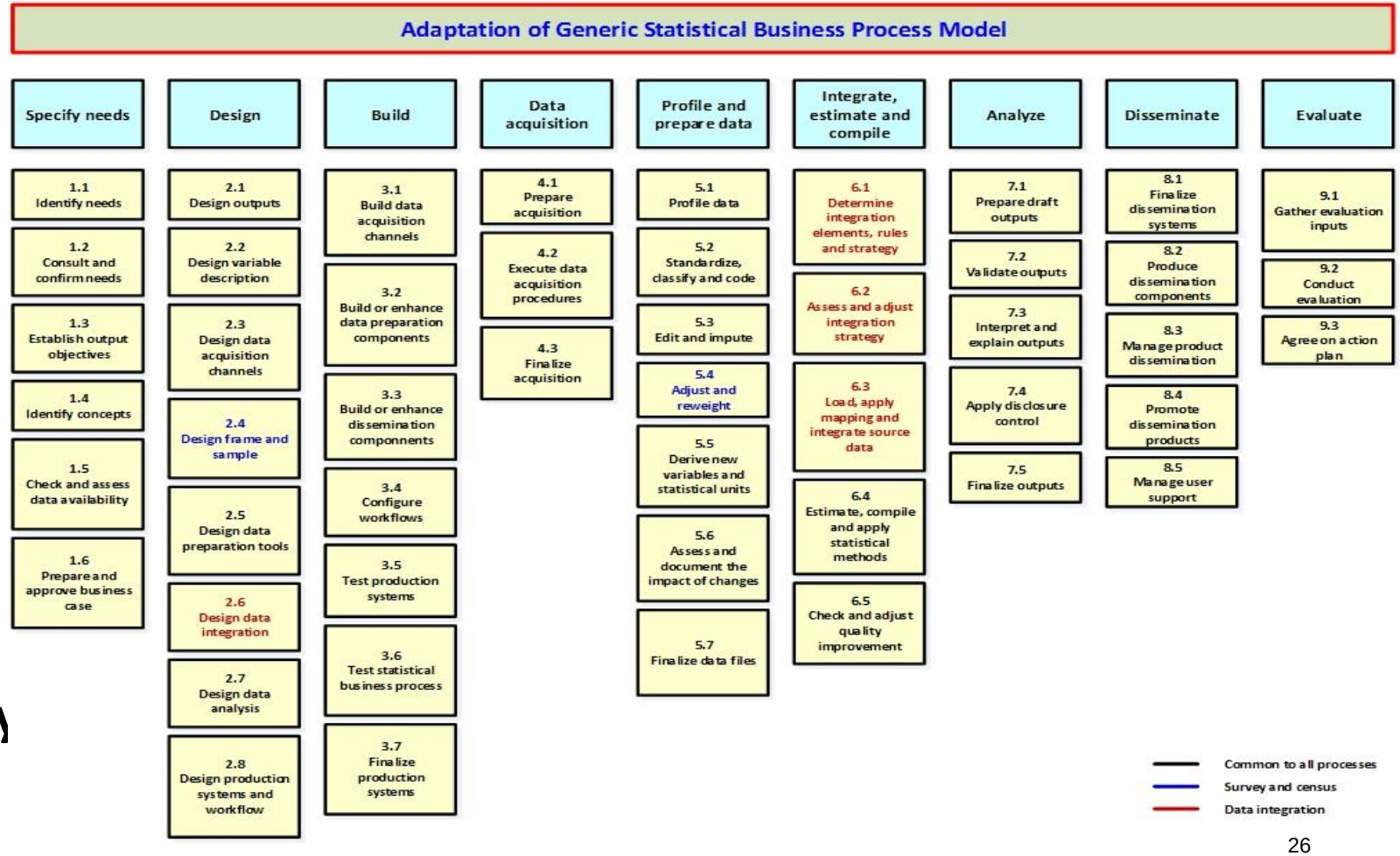
Don’t forget about data ethics!

Even if quality standards are met, the following ethical concerns must be kept in mind when acquiring data:

- The department or data steward’s governing policies
- The reputation of the data provider
- Perceived violation of the privacy rights of individuals
- Potential damage to the department/data steward’s reputation

Generic Statistical Business Process Model (GSBPM)

- Standard framework to assist with the modernization of processes and the sharing of methods/components
- Integration of data and metadata standards
- Harmonization of computing infrastructures
- Framework for process **quality** assessment/improvement



Responsible Machine Learning Framework

RESPECT FOR PEOPLE

- Value to Canadians
- Prevention of harm
- Fairness
- Accountability

SOUND METHODS

- Quality learning data
- Valid inference
- Generalization error
- Explainability

RESPECT FOR DATA

- Privacy
- Security
- Confidentiality

SOUND APPLICATION

- Transparency
- Reproducibility of process and results

Trustworthy insight from ethically sourced data and algorithms

Is Quality Assurance irrelevant in the age of machine learning?

- We can teach machines to recognize patterns to find the data/images (e.g. chocolate croissant) we want. Does this render QA obsolete?

- Wait…sloths?

Quality Assurance Framework (QAF)

- The most recent version released in April 2017
- Serves as the highest-level governance tool for quality management at Statistics Canada
- Twelve stand-alone chapters, each associated with a quality management theme
  - Quality commitment; sound implementation; confidentiality, privacy and security
  - Input data and data providers; managing resources; data users and stakeholders
  - The six quality dimensions

Quality Guidelines

- Most recent version released in December 2019
- Something for everyone: how to build quality into your work
- Quality assurance for all phases of a statistical process
- Relates quality assurance to the dimensions of quality
- Suggests quality indicators
  - Diagnostic: did this step run properly?
  - Informative: what was the impact of this step, and what is the level of quality of this product?

Data Quality Toolkit

- Data quality assurance practices
- Self-assessment checklists
  - Data producer
    - Producer
    - Metadata
    - Data
  - Data user
    - Producer
    - Metadata
    - data

Data user quality assessment checklist

This is an assessment checklist to be completed by anyone contemplating the use of data produced by another organization. The data user should have an idea of what level of quality is required and which attributes are a priority, if they know what they want to do with the data.

Required fields are marked with an asterisk (*).

Questions about the data producer	Answer	Hyperlink to evidence (if applicable)
Q1 Is the data producer a government department, ministry or agency; academic or research institution; private company; or do the data come from another source?	-Select-	
Q2 Do you expect the data producer to still be providing these data one year from now?	-Select-	
Q3 Is the data producer willing to offer you support and consultation about using these data?	-Select-	
Q4 Were documented data quality assurance practices followed in the production of these data?	-Select-	
Q5 Do you perceive the data producer as authoritative on the subject matter of the data and worthy of your trust?	-Select-	

Statistics Canada's Quality Management Tools

Generic Statistical Business Process Model (GSBPM)	Framework for Responsible Machine Learning Processes	Quality Assurance Framework (QAF)	Quality Guidelines	Data Quality Toolkit	Necessity and Proportionality Framework
Standard framework to assist with the modernization of processes and the sharing of methods and components https://statswiki.unece.org/display/GSBPM	Guidance and practical advice on how to responsibly develop these automated processes https://www150.statcan.gc.ca/n1/pub/89-20-0006/892000062021001-eng.htm	Statistics Canada's governance and best practices related to quality management https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.htm	Recommended best practices for each step of the data journey, from design to analysis https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm	Checklist for assessing the usability of a data source: questions about the data producer, data and metadata https://www.statcan.gc.ca/eng/data-quality-toolkit	Balancing societal needs for data insights and the protection of privacy https://www.statcan.gc.ca/en/trust/address

Summary

- What is Quality?
  - Multi-dimensional measure
- How to assess it?
  - Fitness for use depends on users' needs
  - Objective measures are used for assessment
- Where to find references?

Thank you! / Merci !

- **Quality Secretariat, ICMIC**
  - Quality Secretariat team email: statcan.qualitysecretariat-secretariatalaqualite.statcan@statcan.gc.ca
  - Martin Beaulieu: martin-j.beaulieu@statcan.gc.ca