Data sharing solutions at Statistics Canada

G20 DGI-2 Workshop

Promotion of Data Sharing

March 24, 2021

Presented by: Steven Thomas

Statistics Canada

Delivering insight through data for a better Canada

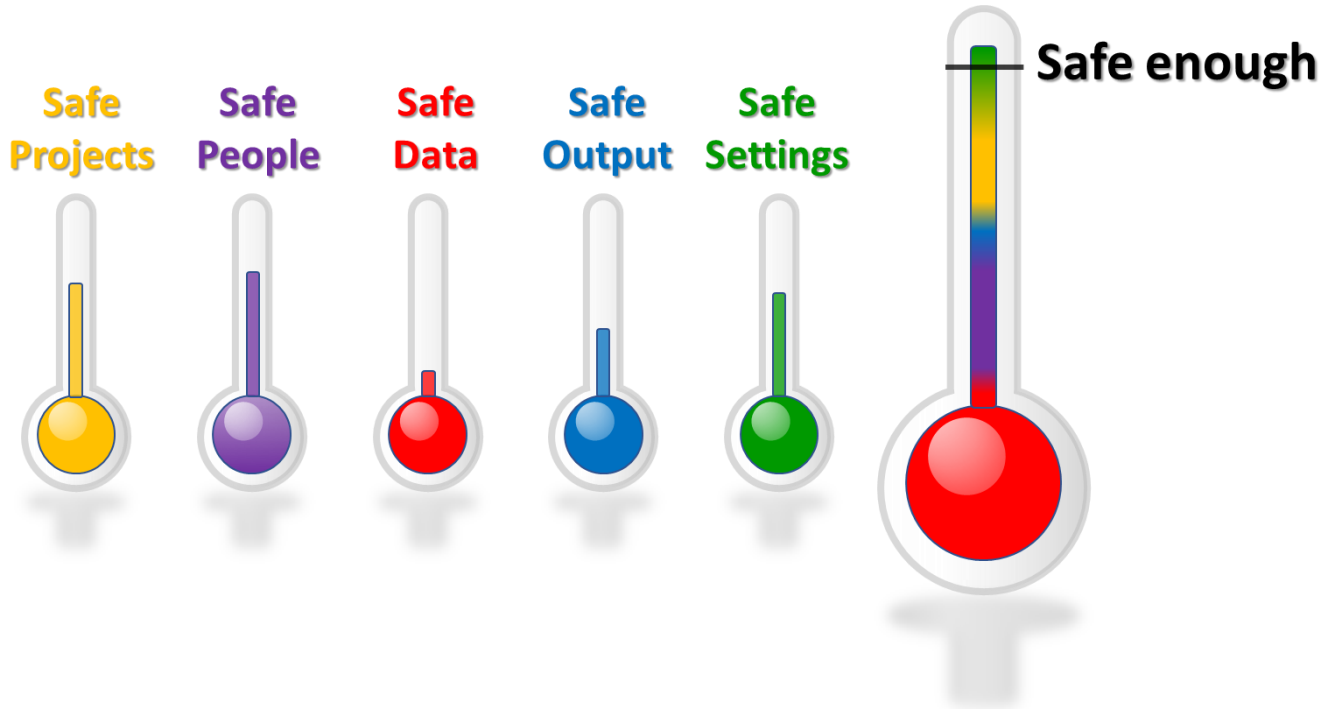
Data sharing - risk and reward

- All data obtained through the Statistics Act must not be disclosed
- At the same time, all data collected through the Act is meant for statistical analysis to help benefit Canadians and must be accessed in some way
- A risk-benefit balance must be determined and this includes allowing researchers direct access to microdata

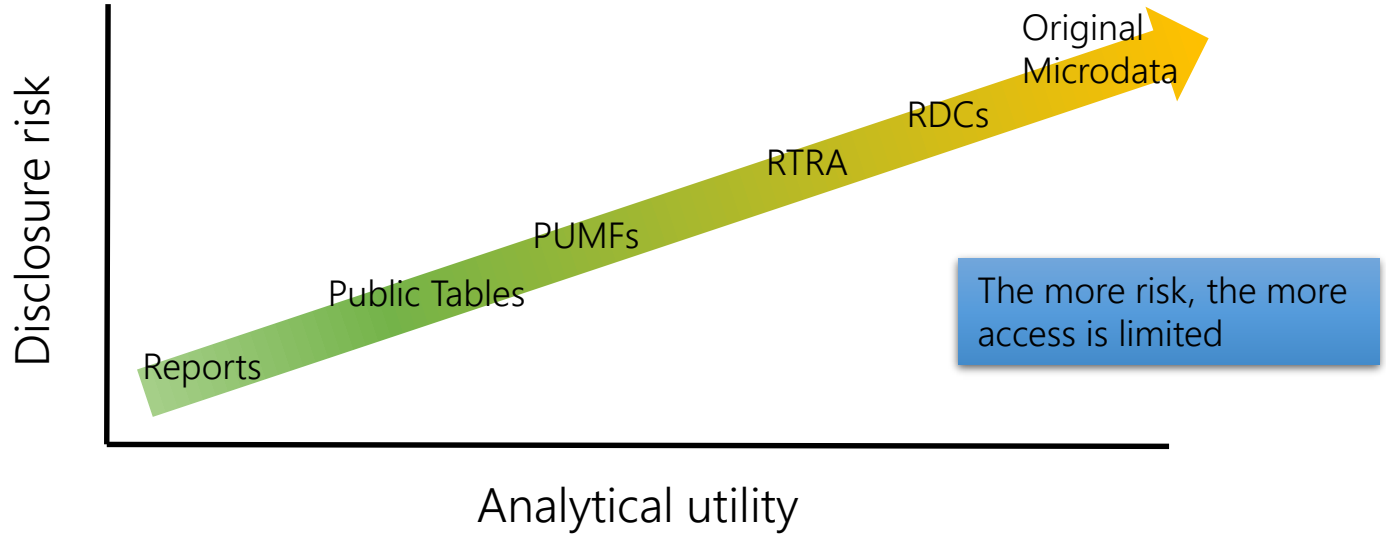
The ‘Five Safes’ Framework

- www.fivesafes.org
- All data access solutions ensure safety through a combination of:
  - Safe Projects
  - Safe People
  - Safe Data
  - Safe Outputs
  - Safe Settings

- Paraphrasing Elliot et al.: ‘safe’ is a goal, not a state → ‘safer’ / ‘safe enough’.



Risk and Utility



Research Data Centres

Facilitate access to microdata files through 32 universities across the country, as well as 3 Federal Research Data Centres

Safe Data

- 182 De-Identified social data files. Access limited to project specific data.

Safe People

- Trusted researcher.

Safe Projects

- Project approval process through institutional review of projects

Safe Output

- Output vetted by StatCan analyst through application of survey specific output vetting rules

Safe Setting

- Microdata remains located in a central and secure location.

- Future work on expanding access to economic data and virtual access through cloud architecture

Public-Use Microdata Files (PUMFs)

- ‘Open’ data solution with unlimited access to 142 anonymized datasets

Safe Data

- Anonymized data. Anonymization processes include de-identification, sub-sampling, aggregation, random adjustments, suppression – lower utility

Safe People

- Subscription based access for institutions requiring microdata access
- Also available through post-secondary institutions

No use of other safeguards

The Real Time Remote Access (RTRA*) system

- Access to over 100 different datasets for over 40 different statistical programs
  - Health, Labour, Income, Diversity, Health Access, Immigration, Education, Cannabis, Spending

Safe Data

- Only specific de-identified datasets are available

Safe People

- The researcher or project team must be associated with a government department, non-profit organization, or an academic institution

Safe Projects

- RTRA users can calculate frequencies, means, percentiles, percent distribution, proportions, ratios, and shares
- Individuals requesting analytical products beyond descriptive and cross-tabular output may wish to explore other options for data access.

Safe Output

- Limited number of output with controlled rounding on requested statistics

Safe Setting

- Microdata remains located in a central and secure location. Researchers do not gain direct access to the microdata and cannot view the content of the microdata file

* More information can be found at https://www.statcan.gc.ca/eng/rtra/rtra

Virtual Data Lab (VDL) Project (New)

Extension of the RDC access model. Developing trust in research partners

Safe People

- Access limited trusted research partners. Shared-risk framework. Partner liability and accountability

Safe Data

- De-Identified Microdata. Admin, integrated and business microdata

Safe Projects

- Institutional review of projects

Safe Output

- Output is vetted against survey specific rules

Safe Settings

- 24/7 Secured Cloud Infrastructure with Physical and IT monitoring

A series of pilots are currently underway with key partners to better improve and strengthen the VDL. Full implementation is scheduled for Fall 2021.

Synthetic Analytical Data – In Development

- Most promising open data option
  - Unlimited access to anonymous microdata
- Terminology and nomenclature
  - Not to be confused with purely random data – Dummy Data
- Methods and tools
  - How to create analytically valuable data without risking disclosure
- 2 cases from StatCan of releasing synthetic analytical data using SynthPop (R package)
- Utility, risk and cost concerns. Can it replace real data access?

Concluding Remarks

- All microdata access solutions require a consideration of the 5 safeguards that allow safe access
- The most sensitive information requires the most protection and use of safety measures
- Open data requires many safety measures to the data itself
- Synthetic Data is an open data option with risk and utility concerns

Thank You! Merci!

For more information please contact / Pour plus de renseignements veuillez contacter

Steven.Thomas@Canada.ca

Views expressed are those of the presenter and may not be those of Statistics Canada

Les opinions exprimées sont celles du présentateur et peuvent ne pas être celles de Statistique Canada