Presentation outline

- BR quality in context
  - What makes for a quality register?
  - Refresher on Statistics Canada’s BR and its sources and processes

- Measurement and mitigation of frame errors
  - Indicators of under-coverage, over-coverage
  - Traditional and new methods of quality assurance and control

- Future challenges and opportunities for BR Quality
Business Register Quality in Context
The BR as a central statistical frame

- The BR is **central** to the **wide array** of economic statistics produced by Statistics Canada
  - Gross Domestic Product
  - Employment statistics (# of employees, wages, work hours)
  - Industrial sales and production data
  - Producer prices
  - Industrial data
  - Business demography and related topics (business financing, success/failure)
  - Many, many other examples

- A **myriad of programs** use and rely on the BR
  - Industrial survey programs
  - Economy-wide survey programs
  - Data integration projects
  - Research and Data Analysis Programs

- It is often used as a **sampling frame**, providing frame listings, and data for industry and size stratification
- Increasingly, it is also used for **allocation** using revenue and employment data and data integration projects
How do we define success in contributing to quality statistics?

<table>
<thead>
<tr>
<th>Dimensions of Quality (Statistics Canada QA framework)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
<td>• BR data content, concepts and classifications are a basis for meaningful statistics</td>
</tr>
</tbody>
</table>
| **Accuracy**                                            | • Listings of active businesses provide good coverage of target populations  
• Businesses attributes needed for classification, stratification and allocation are reliable (i.e. industry codes, employee accounts, revenue values, etc) |
| **Timeliness**                                          | • Point-in-time snapshots align with reference periods prescribed by statistical programs and BR is maintained evergreen |
| **Punctuality**                                          | • Processing and delivery of frame data to statistical programs corresponds to their production schedules with the most up to date information |
| **Coherence**                                           | • Concepts and classifications are standardized  
• Statistical programs use and update the BR data according to standard procedures and processes |
| **Interpretability**                                    | • Up-to-date metadata provided to explain statistics incorporate concepts, definitions and methods of the BR |
| **Accessibility**                                        | • Business demographic data compiled from the BR are available to users inside and outside Statistics Canada |
Basic pre-requisites for a quality statistical register

- **Good, timely data**
  - Administrative data; survey feedback information; info collected by profilers/analysts

- **A good team (or teams) with multi-disciplinary participation**
  - Profilers, economic statistics analysts, methodologists, national accountants, IT professionals

- **Collaborative tools**

- **Standardized concepts classifications**

- **Robust data processing and staff editing/verification procedures (and staff training)**

- **Reliable and sufficient IT infrastructure**

- **Good governance**
Statistics Canada’s BR: Some background

- **6 million active enterprises**, all created and updated on the basis of **tax data**
  - Weekly data transfers from the tax agency

- **3.2 million** are **registered** with the tax agency’s business taxation programs, managed through a common tax identifier – **the Business Number (BN)**
  - For BN-based enterprises, we use **corporate income tax returns, payroll deductions, and sales tax remittances** to update key frame variables

- **2.8 million** enterprises on the BR are **sole proprietors or partners** that have **no BN** (not required if no employees or no sales taxes to pay);
  - these are maintained using personal income tax returns

- **Quality assurance of enterprise-level data** focuses on the **optimal and efficient use of tax information**, which in general accurately portrays the revenues, # of employees, and other frame variables
  - Some challenges to overcome (e.g. industries)
## Distribution of enterprises

<table>
<thead>
<tr>
<th>Business Type</th>
<th>Number of Enterprises</th>
<th>Generated Revenue</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>6 million (95%)</td>
<td>44%</td>
<td>56%</td>
</tr>
<tr>
<td>Profiled Large and Complex</td>
<td>22,000 (0.4%)</td>
<td>53%</td>
<td>41%</td>
</tr>
<tr>
<td>Administrative Complex</td>
<td>260,000 (4.6%)</td>
<td>3%</td>
<td>3%</td>
</tr>
</tbody>
</table>
22,000 profiled enterprises (53% of revenues)

Frame statistical units/structures:

- Legal entities
  - Parent Enterprise
  - Companies (subsidiaries)

- Operating entities (non-legal)
  - Establishments (address sites with employees and cost AND revenue accounting)
  - Locations (address sites with employees and cost OR revenue accounting)

...the work of profilers:

- Review and adjust enterprise ownership/consolidation data

...and the need for quality:

- Drives survey sampling and targeting of questionnaires within complex organizations
- Fundamental to how statistics portray the distribution of economic activity to industries, provinces and local areas. (As frame attributes are used for sample stratification, domain estimation and ratio allocation.)
Measurement and mitigation of frame errors
How do we measure quality and maintain/improve accordingly?

- Various components work together

- 2-pronged strategy:

  1. Invest in independent and quality tracking/measurement activities, conscious of benefits vs costs

  2. Fully leverage existing data, taking advantage of opportunities to find anomalies through data confrontation/comparison
Frame under-coverage

- **Overall** under-coverage is not a major concern and not specifically measured
  
  - Registrations with the tax agency and use of remittances for income taxes, sales taxes and payroll deductions provides a reliable basis for identifying newly-active businesses
  - Business activities not reported for tax purposes are generally not part of the BR’s target universe

- Under-coverage *within industries* can happen, due to:
  
  - **Missing industry codes** for some smaller businesses
  - **Misclassification** (measured by QA survey and other processes)
  - Complex profiles that do not fully and correctly elaborate the operating units engaged in different industries (specific quality control mechanisms)
Frame over-coverage

- Can result from **time lags in the de-activation of units**, because:
  - Tax remittance accounts do not need to be specifically closed upon termination of business activities and income taxes can be reported in arrears
  - BR has wait periods once remittances have stopped prior to deactivating

- **Industry misclassification** causes under-coverage in one industry and over-coverage in another
Unclassified units (contributes to industry under-coverage)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>1,297,339</td>
<td>779,412</td>
<td>540,423</td>
<td>531,949</td>
<td>531,969</td>
<td>558,755</td>
</tr>
<tr>
<td>Ent. Population (%)</td>
<td>28.8%</td>
<td>14.5%</td>
<td>9.4%</td>
<td>8.9%</td>
<td>8.8%</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

[Graph showing the decline in counts and percentage]
Results from BR Quality Assurance Program

- Annual sample of 3,600 businesses contacted to provide estimates of:
  - Missed de-activations
  - Misclassification

- For North American Industry Classification System (NAICS) sectors (20 groupings)
Industry misclassification

☑ Error

<table>
<thead>
<tr>
<th>Year</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on Frequency</td>
<td>11.5%</td>
<td>12.4%</td>
<td>13.1%</td>
<td>14.7%</td>
<td>13.1%</td>
</tr>
<tr>
<td>Based on Revenue</td>
<td>0.0020%</td>
<td>0.0014%</td>
<td>0.0021%</td>
<td>0.0026%</td>
<td>0.0012%</td>
</tr>
</tbody>
</table>

☑ Volatility

<table>
<thead>
<tr>
<th>Year</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on Frequency</td>
<td>4.6%</td>
<td>3.3%</td>
<td>4.2%</td>
<td>4.0%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Based on Revenue</td>
<td>0.0004%</td>
<td>0.0002%</td>
<td>0.0002%</td>
<td>0.0002%</td>
<td>0.0002%</td>
</tr>
</tbody>
</table>

☑ Total

<table>
<thead>
<tr>
<th>Year</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on Frequency</td>
<td>16.1%</td>
<td>15.7%</td>
<td>17.3%</td>
<td>18.7%</td>
<td>16.7%</td>
</tr>
<tr>
<td>Based on Revenue</td>
<td>0.0024%</td>
<td>0.0016%</td>
<td>0.0023%</td>
<td>0.0028%</td>
<td>0.0014%</td>
</tr>
</tbody>
</table>
## Industry misclassification

**Average NAICS Error Rate (ANER)**  
Based on Frequency and based on Revenue

<table>
<thead>
<tr>
<th>Sector</th>
<th>Freq</th>
<th>Rev</th>
<th>ANVR</th>
<th>ANMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-Agriculture, Forestry, Fishing and Hunting</td>
<td>3.5%</td>
<td>0.0016%</td>
<td>1.1%</td>
<td>2.4%</td>
</tr>
<tr>
<td>21-Mining and Oil and Gas Extraction</td>
<td>23.0%</td>
<td>0.0012%</td>
<td>4.3%</td>
<td>18.7%</td>
</tr>
<tr>
<td>22-Utilities</td>
<td>15.7%</td>
<td>0.0001%</td>
<td>1.1%</td>
<td>14.6%</td>
</tr>
<tr>
<td>23-Construction</td>
<td>16.5%</td>
<td>0.0010%</td>
<td>1.5%</td>
<td>15.0%</td>
</tr>
<tr>
<td>31-33-Manufacturing</td>
<td>26.5%</td>
<td>0.0007%</td>
<td>5.6%</td>
<td>20.9%</td>
</tr>
<tr>
<td>41-Wholesale Trade</td>
<td>34.2%</td>
<td>0.0010%</td>
<td>10.2%</td>
<td>24.0%</td>
</tr>
<tr>
<td>44-45-Retail Trade</td>
<td>22.4%</td>
<td>0.0011%</td>
<td>6.9%</td>
<td>15.5%</td>
</tr>
<tr>
<td>48-49-Transportation and Warehousing</td>
<td>12.3%</td>
<td>0.0002%</td>
<td>3.7%</td>
<td>8.6%</td>
</tr>
<tr>
<td>51-Information and Cultural Industries</td>
<td>25.4%</td>
<td>0.0040%</td>
<td>4.3%</td>
<td>21.1%</td>
</tr>
<tr>
<td>52-Finance and Insurance</td>
<td>19.5%</td>
<td>0.0022%</td>
<td>1.0%</td>
<td>18.5%</td>
</tr>
<tr>
<td>53-Real Estate and Rental and Leasing</td>
<td>11.9%</td>
<td>0.0013%</td>
<td>3.4%</td>
<td>8.5%</td>
</tr>
<tr>
<td>54-Professional, Scientific and Technical Services</td>
<td>16.4%</td>
<td>0.0008%</td>
<td>4.2%</td>
<td>12.2%</td>
</tr>
<tr>
<td>55-Management of Companies and Enterprises</td>
<td>54.3%</td>
<td>0.0065%</td>
<td>5.1%</td>
<td>49.2%</td>
</tr>
<tr>
<td>56-Administrative and Support, Waste Management and Remediation Services</td>
<td>33.2%</td>
<td>0.0052%</td>
<td>5.9%</td>
<td>27.3%</td>
</tr>
<tr>
<td>61-Educational Services</td>
<td>17.0%</td>
<td>0.0030%</td>
<td>2.9%</td>
<td>14.1%</td>
</tr>
<tr>
<td>62-Health Care and Social Assistance</td>
<td>6.5%</td>
<td>0.0005%</td>
<td>1.6%</td>
<td>4.9%</td>
</tr>
<tr>
<td>71-Arts, Entertainment and Recreation</td>
<td>23.5%</td>
<td>0.0043%</td>
<td>4.9%</td>
<td>18.6%</td>
</tr>
<tr>
<td>72-Accommodation and Food Services</td>
<td>6.7%</td>
<td>0.0006%</td>
<td>2.1%</td>
<td>4.6%</td>
</tr>
<tr>
<td>81-Other Services (except Public Administration)</td>
<td>13.0%</td>
<td>0.0071%</td>
<td>4.0%</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>16.7%</td>
<td>0.0014%</td>
<td>3.6%</td>
<td>13.1%</td>
</tr>
</tbody>
</table>
New method for assessing NAICS codes

- QA survey sample size limitations

- We now have multiple sources for NAICS coding that can be used to ‘confirm’ one another:
  1. Tax agency registration
  2. Self-coding on annual income tax forms (for corporations and sole proprietors/partners)
  3. Verified/updated codes from ‘survey feedback’ (online questionnaire component)
  4. Updates by profilers and analysts
Composite rating of NAICS quality based on vintage of current value, number of confirmations and sector accuracy (QAP)

By counts

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13%</td>
<td>76%</td>
<td>11%</td>
</tr>
</tbody>
</table>

By revenue

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65%</td>
<td>25%</td>
<td>10%</td>
</tr>
</tbody>
</table>
Frame duplication: on-going risks

1. (Unincorporated) sole proprietors sometimes have tax registrations for sales taxes or payroll remittances, but we don’t always know the links
   - Can result in duplicate records on the BR; risks are attenuated for industries with higher concentration of small firms (e.g. agriculture)
   - Working intensively to improve record matching techniques, using better name parsing and introducing geo-locating methods
   - Also introducing collection of tax registration identifiers on survey questionnaires

2. Increasingly, businesses appear to be processing tax remittances through separately-registered tax accounts, sometimes through separately contracted firms
   - Profilers must take care not to double-count businesses, employees and revenues
Maintaining quality of profiler and analyst frame updates

- **Training** curriculum for updater/profiler; required for system privileges

- Close **collaboration** and consultation across **profiling teams and statistical subject-matter programs**

- **Quality Control module** recently implemented (audit recommendation) to sample and verify integrity of updates of staff from across Statistics Canada and follow-up as required
Other QA methods for complex profiles

- **Critical flag** prioritizes profiles for review (prior to December frame snapshot) and used in many annual surveys

- Each complex has a **composite integrity score**, based on:
  1. Additivity of (hierarchical) employment and revenue values (i.e. *vertical coherence*)
  2. Recency of last profile update
  3. Counts and types of ‘messages’ issued by analysts requesting reviews or other signals of possible changes
  4. Consistency of provincial employment distributions with annual summary of earnings and tax deductions sent to employees (T4 slips)

- Additionally, formalized methods are in place to compare BR frame variables (employment, revenues) with a variety of survey and administrative sources (horizontal coherence)
Current state in summary

- **Internal audit** of BR QA practices (2016) was **positive**

- **Frame coverage issues within industries are minor** in the context of economic activity; employees and revenues are being classified to the industries and provinces where they belong.

- **Some issues** remain with “micro businesses” that have no identified links to the revenue agency’s business registrations and for which the precise moment of termination of business activity can be difficult to discern from the tax data.
  - These issues will need to be mitigated as we build the longitudinal entrepreneurship database.

- As the need for **local statistics** grows, we are increasingly challenged to find **more timely and efficient ways of elaborating complex enterprise statistical locations**.
Future challenges and opportunities
The BR in a newly envisioned & modernized statistical system

- Newly-launched Statistics Canada modernization initiative challenges us to move to admin/big data sources wherever possible (instead of surveys) and ensure collection and processing methods keep pace with technology and advancements.

- Heightens the role of central registers as hubs for data linkage; requiring quality control for reconciliation of data records with register statistical units.

- We have begun working towards an integrated system of core registers of buildings, people and businesses; record linkage and frame quality assurance are key project components.

- We are also looking for alternative methods and sources for complex enterprise profiling, such as outward facing web portal or usage of third party data; impacts on data quality will need to be taken into consideration.