**Description of Data Access, Indicator Generation & Screening – CSDUL, Node 3**

**Description of Data Access** (Downloading of the data, programming, aggregation, etc. - anything GIS related)

*Data acquisition*

FRSR indicators are based on the Enhanced Points of Interest (EPOI) data product from DMTI Spatial, which provides information on various community resources across Canada including containing x/y coordinates and categorical information . These raw datasets were downloaded in point shapefile format via Scholar’s Geoportal. Geographical census boundary datasets to aggregate FRSR counts within were downloaded from Statistics Canada as polygon shapefiles.

*Data preprocessing*

All geospatial datasets were unzipped and organized on a local file system, with qualitative checks applied in QGIS. To account for slight differences in EPOI file delivery (e.g. some years were broken down by province), a single merged canada\_epoi\_20XX.shp was created for each year under consideration (2011, 2012, 2013, 2015).

*Resource count extraction*

Specific resource types from the EPOIs were categorized using Standard Industrial Classification (SIC) codes, which identify specific resource types like education facilities, food stores, recreation facilities, and police and fire stations. A list of SIC codes corresponding to specific FRSRs was developed by the research team, available [here](https://docs.google.com/document/d/11ahpEihyLESNAywg8IPRK8RcLPsziVaz/edit#heading=h.7f9l3ac2qacw).

For each year under consideration, the Python programming language was used to load the national EPOI dataset and census tract boundaries in geodataframe format. A FOR loop was used to extract only the EPOIs corresponding to specific SIC codes (e.g. all points where SIC\_1 == 8221) and count the number of these points within each census tract polygon. After looping through all SIC codes within the provided list, a CSV file and polygon shapefile containing point counts for each FRSR code are saved to disk. Optionally, the subset of individual points for each SIC code can be exported as well for spot-checking purposes.

*Quality control*

After an exhaustive spot-checking process, two filtering mechanisms were implemented to remove redundant or irrelevant points from the EPOIs before counting resources. For any points with exactly the same latitude and longitude coordinates, only one was retained for the census geography counts. This reduced the overrepresentation of educational infrastructure, where potentially dozens of points for a single university or college were counted under SIC 8221. Additionally, a keyword search was used to drop any points with names containing the strings “lotto” or “lottery” from SIC 7999 and “rmt” from SIC 8299.

Shapefiles containing raw point counts and post-filter results have been provided for each year. See [here](https://docs.google.com/document/d/1lycc3JaAVn4PNslXmE6TZal_w3pEEPt2/edit) for a detailed explanation of spot-checking results.

**Spot-Checking Process** (Overview of steps, notes and recommendations from previous document)

The Spot-Checking process involves the cross-verification of the constructed FSFR indicators with geospatial data obtained from Scholar’s GeoPortal through DMTI Spatial. This process involves confirming that data points, which we classified using SIC codes, are correctly matched and aggregated to their corresponding Census Tracts (CTUID). To validate the data points, we download the prepared Shapefiles into ArcGIS Pro for manual inspection. For each FSFR category, which we identify through corresponding SIC codes, we then select, at random, two census tracts per code and perform two independent checks. Within each selected tract, we first visually verify that the number of points identified and the type of facilities assigned to that SIC code align with the indicator values. In doing so, we cross-reference the spatial placement of these facilities using Google Earth historical imagery, referencing the ascribed year of data. With confirmation that the data points correspond to the actual facilities on the ground, we can then confirm the accuracy of these data.

As an example, if we are conducting a check with data from 2015 for SIC code 8221, postsecondary institutions, we would load the Shapefile into ArcGIS Pro and select a Census Tract. If this selection displays two points for SIC code 8221 within this tract, we would then inspect these data points manually to determine if they are indeed two independent facilities, confirm that these facilities are postsecondary institutions, and verify that they fall within the correct Census Tract. Following this, we would inspect these locations on Google Earth, and using historical imagery, review these sites as they appeared in 2015.

This process allows us to assess the accuracy of the data as well as the accuracy and reliability of the spatial joins.

Following point extraction, several cleaning and quality-control measures are implemented to improve indicator accuracy across education and recreation categories specifically. Duplicate entries have been removed, with deduplication based solely on address remaining under consideration, as some sites may host multiple distinct resources, as seen in overcounts for SIC 8221 (postsecondary institutions). Keyword-based filters have also been introduced for SIC 7999, with points containing “lottery” or “lotto” excluded to prevent the conflation of gambling-related services with socially beneficial recreation. However, we acknowledge that data such as these may be of use to other researchers, so further refinement may be needed. Additional refinement is also being explored for SIC 8299, where misclassified therapeutic clinics (e.g., massage or RMT providers) may inflate counts. Manual review may be needed to ensure valid count retention if keyword filters are introduced. Additionally, unique to the province of Quebec, SIC 8222 has been retained as having educational resources when aggregated according to the FSFR paper’s classifications. Due to this regional specificity, flagging or excluding this SIC from national composite indices may be of interest.