UNECE High-level Group for the
Modernisation of Official Statistics

Synthetic Data for Administrative Sources

This business case was prepared by Statistics Canada in collaboration with the Blue Sky Thinking Network Core Group and submitted to the HLG-MOS for approval.

Type of Activity			
☐	New project	☒	New activity
☐	Extension of existing project	☐	Extension of existing activity
Projects are undertaken by separate project teams. Projects are expected to produce a significant contribution to achieving the HLG-MOS vision		*Activities are undertaken by Modernisation Groups. These activities produce smaller, more detailed outputs to help achieve the HLG-MOS vision*	

Purpose

Modern statistical organizations operate in an environment that continues to call for more open sharing of data, expertise and best practices both across their internal business processes as well as with external partners. The unequivocal commitment that NSOs make to trusted data protection along with increasing demands for access to timely data from an ever-growing body of new, large and complex data sources highlight the need for a modernized approach to data sharing.

Synthetic data sets have become a desirable alternative to public use microdata files because they may be capable of retaining a more significant proportion of analytic value while protecting confidentiality. Use of high-quality synthetic data can also improve data integrity through such statistical processes as testing new systems and data modeling, particularly for large, administrative data sources and privately-held big data (together referred to as big data below). Current techniques are time and resource intensive and as a result, have not been broadly scaled and cannot address the unique requirements that come with current complex big data sources. A modern, interoperable approach to creating and using synthetic data sets would yield efficiencies in timeliness, quality and coherence.

As big data's contribution to official statistics continues to grow, a new approach is needed within the data ecosystem to accelerate and streamline its integration into statistical production using interoperable, privacy-preserving approaches. Testing an effective approach to creating synthetic data sets from big data sources has high value for both internal use (IT system development, analysis, modeling) as well as making information available to external partners. Several uses cases can be envisaged but experience in what works and when can only be gained through hands-on experience.

The goal of this activity is to run an experiment to generate and test synthetic data with proper format and structure from a large administrative data set using modern approaches including open source tools. The project is expected to generate significant opportunities for learning and sharing across the statistical community as well as with external partners and stakeholders.

Description of the activity

The activity is divided into three phases:

Phase 1. Create a relevant experiment charter including specific statistical use-case(s).
The first step is to investigate and document statistical use cases that could benefit from synthetic data and establish parameters for the experiment. Statistics Canada has created a draft experiment charter to initiate discussions. Participating experts will be invited to validate the experiment approach and collaborate on identifying a type of data set relevant to a wide international audience.

Phase 2. Execute the experiment
The second step is to execute the experiment using the scientific approach to test the feasibility of using open source tools to generate synthetic data from a large administrative data set. This step would again leverage the expertise in NSOs as well as external partners and stakeholders (academics, private sector) in the area of data infrastructure innovation. Some first contacts with potential partners have already been established.

Phase 3. Share lessons learned and map out path to operationalization in production environments
During this phase, the team will review results of the experiment, document and broadly share lessons learned. In addition, it will build a roadmap with next steps to operationalize modern synthetic data into statistical production.

Alternatives considered

Directing this proposal to an existing Modernisation Group was considered but there is no obvious fit. Results of this activity can be leveraged by the Machine Learning project as preparing administrative data in the proper format and structure is an important precursor for ML projects. The project can be scaled up to include more application areas and uses cases.

How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?

This activity will advance research and adoption of practical approaches to harvest the value of data synthesis in the statistical production process, particularly for testing, modelling and sharing data with external partners. The experiment will build capacity in the area of data infrastructure innovation and inform future work on the use of open source tools.

Proposed start and end dates

Start: *January 2020*	**End:** *December 2020*