A New Model for International Privacy Preserving Data Sharing Across National Statistical Organizations

Curtis Mitchell[1], Ben Santos[2], Massimo De Cubellis[3], Angela Pappagallo[3], Sean Lovell[4], Ian Munoz[1], Kate McCall-Kiley[1], Luke Keller[1], Stephanie Studds[1], Beau Houser[1], Monique Eleby[1], Ronald Jansen[4]

United States Census Bureau[1], Statistics Canada[2], Italian National Institute of Statistics[3], United Nations Statistics Division[4]

Abstract

National Statistical Organizations (NSOs) such as the U.S. Census Bureau have competing mandates to both maintain the privacy of individuals and organizations whose data they collect and to provide broad access to that data. This often means that access to NSO data is highly restricted and regimented in order to minimize the risks of privacy violations. The emergence of privacy-enhancing technologies, or PETs, are creating opportunities to reconsider how NSOs can make data more available while preserving the privacy of subjects within the data. Here we report on an ongoing collaboration between the Census Bureau, Statistics Canada (StatCan), and the Italian National Institute of Statistics (Istat) in conjunction with the United Nations Privacy-Enhancing Technologies Lab (UN PET Lab). The project involves using the open-source platform PySyft and establishing the digital infrastructure necessary so that nodes hosted by the Census Bureau, StatCan, and Istat, facilitated by a network gateway hosted by the UN PET Lab, can perform a join on synthetic data. This will allow for testing and building confidence before attempting joins on actual private data. We believe this project will be an important milestone towards enabling international, privacy-preserving data science between government agencies.

1. Introduction

National Statistical Organizations (NSOs) collect data about their nation's population and economy, and are tasked with making that data broadly available while also maintaining the privacy of individuals and organizations reflected in that data. Currently, international data exchanges between NSOs consist of a lengthy but necessary series of steps to verify the identity of researchers and validate their use-cases before any analysis of NSO data can begin. Even after the data exchange has begun, this exercise frequently involves numerous additional processes such as tracking where data is flowing and who has access to that data. The adoption of privacy-enhancing technologies (PETs) offers a means to ease the administrative burden of several parts of this data exchange operation. This pilot study is intended to achieve multiple objectives, primarily around determining the feasibility of a joint computation on data between two or more national statistical agencies using PETs.

The UN PET Lab is an initiative established by the UN Committee of Experts on Big Data and Data Science for Official Statistics. The intent is that the lab will consist of "A series of active proofs-of-concept and pilot projects focused on the evaluation of PETs for real-world use cases in the official statistics community" (1).

The Census Bureau is responsible for the creation and maintenance of data regarding the population and economy of the United States. In conjunction with the Office of the Chief Information Officer (OCIO) within the Census Bureau, xD (2) is the team exploring use cases and pilot projects for potential PETs applications both within the US federal government and with external public- and private-sector partners. In charge of this project is a cohort of technology professionals within the xD team, the Emerging Technology Fellows (ETFs), a group of experienced technologists performing a term-limited tour of public duty through the fellowship program.

Statistics Canada (StatCan) acts as Canada's NSO and provides statistical information and data about Canada's economy, society, and environment to help Canadians to function effectively as citizens and decision makers. The agency already has rigorous measures in place to preserve privacy and confidentiality in the modern digital era and this commitment remains the highest priority. In the Data Science and Innovation division at StatCan, a dedicated team of researchers has been exploring emerging PETs and their potential to address the privacy preservation needs for highly sensitive information in various scenarios; a short list of related projects can be found in (3). In addition to alternative storage options, PETs will allow the agency to adopt and implement remote and delegated computing on encrypted data, benefit from potential multi-party computation opportunities and derive insights from distributed and inaccessible data (4). For this reason, engaging with other NSOs and vendors working on PETs and emerging technologies is essential, and the UN PETLab has demonstrated to be the right venue for sharing findings and strengthening collaborations like the one described in this paper.

The Italian National Institute of Statistics, (Istat) is the NSO responsible for population and economic censuses as well as social, environmental, and economic surveys and analyses within Italy. Istat has been engaged in research on privacy-enhancing technologies for several years. For example, the Institute took part in a number of experimental projects in the context of the PETs Lab and has activated a Trusted Smart Surveys center which includes privacy-related requirements. Furthermore, PETs are transversal techniques that can be applied in different domains within official statistics. In this regard, Istat frequently needs to integrate its own data with data held by other public administrations for the sake of producing official statistics. Because of these requirements, Istat was very interested in participating in this pilot project with the UN PET Lab and other NSOs.

2. Project objectives

This pilot study consists of multiple objectives to determine the feasibility of a joint computation on data between two or more national statistical agencies using privacy-enhancing technologies (PETs). Additional objectives include:

Protecting Data Privacy: The protection of data privacy is extremely important for entities like NSOs who not only have legal obligations to protect respondent privacy but rely on their reputation as strong data stewards to foster public cooperation with their surveys and censuses. Because the pilot only uses open, harmonized trade data from UN Comtrade, it offers the opportunity to explore the extent to which the technology delivers on its promise to protect data privacy.

System Security: Because of the sensitivity of NSO data, another key objective of the pilot is exploring security requirements for creating a joint computation. The pilot needed to work in such a way that it enabled connections for Census users without crossing any boundaries into Census Bureau systems that store and process confidential information.

Evaluating Open-Source Software: This pilot utilizes a free and open-source platform called PySyft(5) designed and maintained by OpenMined (6). The objective is to evaluate the usage of such open-source software as it has numerous benefits over traditional software, including a free license and a reviewable and auditable codebase.

Evaluating PETs: The UN PET Lab Pilot is allowing NSOs to test several PETs through PySyft, many of which are still early in development. PySyft itself is in version 0.8. PySyft's data flows follow a framework called structured transparency to enable the addition of PETs that fulfill an input privacy function, such as secure multiparty computation (SMPC), as well as PETs that fulfill an output privacy function, such as differential privacy (DP), to create an end-to-end data privacy solution.

Future Capabilities: In addition to including several data privacy controls, PySyft is also a federated learning framework that supports more complex capabilities in the form of machine learning algorithms that can be run across multiple remote datasets simultaneously. Questions remain about the ability of software-driven input privacy methods such as SMPC and homomorphic encryption to execute resource-intensive queries such as machine learning algorithms. However, having the means to deploy these types of solutions now will enable usage of these forward-thinking capabilities in a production setting sooner than later.

Deploying PETs on Traditional Hardware: With a few exceptions, many PETs can be used without specialized hardware requirements. This has allowed the team to navigate traditional IT governance structures using infrastructure that is readily available for use as opposed to going through a lengthy acquisitions process such as, in the case of the Census Bureau, assuming a vendor’s solution has approval by the Federal Risk and Authorization Management Program (FedRAMP)(7).

Interpreting Policy: An additional goal is to understand whether the existing policy creates an opportunity for technologies such as the PySyft framework that can achieve or facilitate a reduction in operational burden for participants to share or deliver data products in addition to improving or reducing the envelope or perimeter of risk. Part of this process would be a discussion regarding how existing policies may need reevaluating as these technologies mature and are deployed in production settings.

3. Technical Overview

The primary tool used for this project is PySyft, a Python-based API for the “Syft” framework. Syft is an open-source tool that enables external data scientists to submit their code to data owners who control access to one or more remote, privately held datasets. When combined with the use of PETs such as differential privacy, Syft can enable remote, privacy-preserving data science.

3.1 Syft Role Definitions

There are several roles and responsibilities needed to utilize the Syft network:

Data Owner – A user who maintains access to private data and approves or denies code submissions to be run on that data as well as approving or denying the sharing of results from that a code run

Data Scientist – A user submitting code that is approved or denied to run on a private dataset

Ambassador – An optional user who can also review code submissions and approve or deny access to a high-side domain

Domain Server – A server that has the ability to host data and provides the ability to review and run submitted code on that data

- Low-Side Domain – a domain server that is public and discoverable
- High-Side Domain – a domain server that is private and gated

Gateway Server – A server that acts as a bridge between domains and data scientists, data owners, and ambassadors to enable collaboration and data analysis between multiple parties

3.2 PySyft Data Flow

The data flow goes through several steps to prevent a data scientist from having direct access to sensitive data and provide sufficient opportunities for code they submit to be reviewed. It's worth noting that is process does not necessarily have to be done instantaneously, by, for example, all parties being online and available at the same time.

The steps are:

1. The data scientist logs onto the Gateway node and sees one or more attached low-side domains.
2. Using mock data in the low-side domains that has the same schema and shape (but different values) as the actual private data, the data scientist crafts a module of Python code they hope to run on the private data set(s).
3. (Optional) The data scientist submits the code to the Ambassador node on the low-side domain. The ambassador may reject the code sample and request changes, or approve the request and forward it to the high-side gateway.
4. One or more data owners receive the code submission, review it, and approve or deny the request for the code to run on private data.
5. The code is run on private data (optionally on a secure enclave).
6. The data owners review the resulting output of the code run and either approve or deny the sharing of that result.
7. (Optional) If the code result is approved by the data owners, the result is sent back to the high-side gateway for review by the ambassador. If approved the ambassador passes the result onto the low-side gateway.
8. The code result is returned to the data scientist.

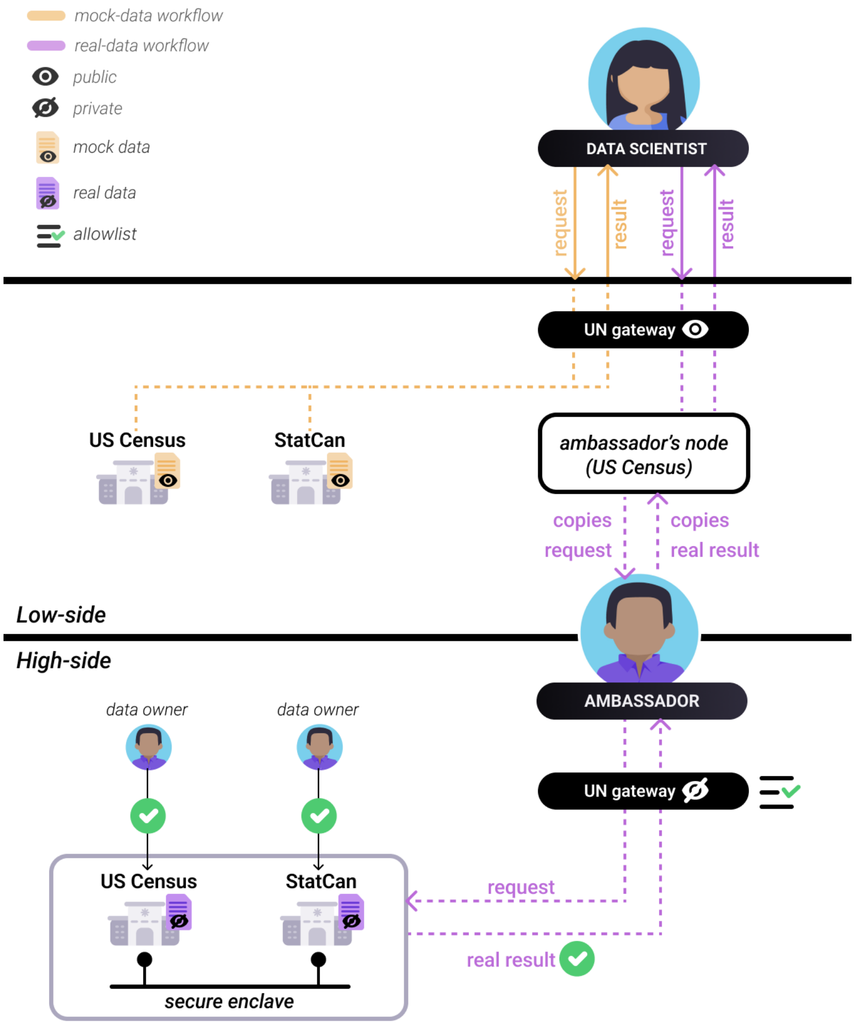

Figure 1 - Dataflow in PySyft between Data Scientist and NSOs

3.3 Deployment Environment

To create an environment to deploy Syft and connect to the UN PET Lab, the Census Bureau team initially established a secure Azure environment. The team later determined that a simpler deployment configuration could be achieved with lower maintenance costs and requirements using Cloud.gov (8), a government-focused service based on the Cloud Foundry platform. By launching a PySyft server (and Jupyter notebook server to interact with that server) on Cloud.gov, the xD/Census team was able to establish its domain on the Syft network and connect to the UN gateway to enable connections to other NSO nodes with minimal overheads and maintenance requirements on a system that is separated from the general Census network.

For Istat's deployment, the team initially set up a Microsoft Azure virtual machine with an exposed IP address in a secure environment. With the support of the OpenMined team, they were able to install and configure a PySyft node to connect to the UN gateway. This enabled the Istat team to establish a connection to the PET Lab network, facilitating secure communication with the other NSOs involved in the project.

StatCan leveraged the Shared Services Canada's (SSC) Science Program (9) to create an Azure subscription on their cloud, serving as a sandbox environment for unclassified data. Within this subscription, the team deployed an Azure Virtual Machine and, with the assistance of OpenMined, set it up as StatCan's instance of a PySyft lowside domain. Statcan then established a connection to the UN gateway while OpenMined hosted the highside domain for StatCan.

3.4 Dataset Description

For the purposes of illustrating the technical feasibility of the PySyft platform to perform cross-border data analysis, the team used a subset of data of artificially-generated census records used as part of the "Freely Extensible Biomedical Record Linkage" (FEBRL) project, an open source data processing and analysis platform (10). These datasets are included as part of the Python-based Record Linkage Toolkit library (11) used to perform the data join in this project. The datasets consisted of 819 rows of fictitious census records.

4. Results

In a demonstration in February 2024, StatCan and Istat were able to perform a successful join on their data, followed by a successful join between StatCan and the Census Bureau datasets in May 2024 (12). During the demonstration join, a data scientist was able to successfully submit Python code. Following review of the code submission by an ambassador, data owners representing both NSOs reviewed and approved the code run and result, which was shared back to the data scientist.

4.1 Data Join Technique and Shared Result

To illustrate the ability to perform a data join between two private datasets, the Python code submitted by the data scientist utilized the Record Linkage Toolkit library to compare individual census records across both datasets hosted by the NSO domains. A record was specified as a match when an entry in both datasets had the following fields in common:

- given name
- surname
- date of birth
- suburb
- state
- address

The value reviewed by the data owners and ambassador and returned to the data scientist was the total count of matching records present in both NSO datasets, a count of 813 shared records in the case of the US Census-StatCan join.

5. Discussion and Next Steps

This project is intended to demonstrate the feasibility of using PETs, and PySyft specifically, to enable joint computations between various NSOs. Now that a basic join has been demonstrated, the Census

Bureau team are in discussions with OpenMined, StatCan, and Mexico's NSO, the National Institute of Statistics and Geography (INEGI), to perform a join modeling North American trade data represented by synthetic micro trade data generated by each respective NSO.

The initial data join described is this paper is considered a "Phase 1" of this project for the Census Bureau. Phase 2 will consist of utilizing synthetic data based on actual private data in 2025. This will allow the teams to begin working towards a more robust system architecture before attempting a join on actual private data, which is Phase 3 and the final phase of the pilot project.

PETs are still evolving and maturing at different rates, and this varies even by technique. This requires a thoughtful approach to their testing and use while also seeking to realize their strategic benefit. Reports like the UN PET Lab Guide (13), the US National Strategy to Advance Privacy-Preserving Data Sharing and Analytics (14), and a review of privacy-enhancing technologies published by the Office of the Privacy Commissioner of Canada (15) highlight the potential that these techniques bring and the need for pilot use cases that demonstrate not only the feasibility of the technologies themselves but the policy, legal, and sociotechnical aspects that will contribute to the health of a future ecosystem wherein these technologies are employed.

Acknowledgements

Ron Jarmin, US Census Bureau

Christa Jones, US Census Bureau

Luis Cano and the Office of the Chief Information Officer, US Census Bureau

Nick Orsini and the Economic Directorate, US Census Bureau

Policy Coordination Office, US Census Bureau

James McCoy, US Census Bureau

Michael Castro, US Census Bureau

The entire OpenMined team

References

1. https://unstats.un.org/bigdata/task-teams/privacy/index.cshtml
2. https://xd.gov/
3. Privacy Enhancing Technologies at Statistics Canada. Proceedings of the Survey Methods Section, SSC Annual Meeting, May 2022. https://ssc.ca/sites/default/files/imce/dugdale_molladavoudi_ssc2022.pdf
4. Government of Canada, (2021, December 15). Data Science at Statistics Canada. Statistics Canada. https://www.statcan.gc.ca/en/data-science/stat
5. OpenMined/PySyft: Perform data science on data that remains in someone else's server (github.com)
6. https://openmined.org/
7. https://www.fedramp.gov/
8. Home | cloud.gov
9. https://wiki.science.cloud-nuage.canada.ca/en/public
10. https://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html
11. Python Record Linkage Toolkit Documentation — Python Record Linkage Toolkit 0.15 documentation
12. YouTube Recording, Census – UN – StatCan Data Join Using PySyft, May 23, 2024: https://www.youtube.com/watch?v=cq8TdMysA3Y
13. https://unstats.un.org/bigdata/task-teams/privacy/guide/
14. https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf
15. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2017/pet_201711