Attachment 1 – Project Topic

Future Directions for Data Collection Frames and Novel Estimation Strategies

Key Objective

This project seeks new adaptations to the rapidly changing landscape of survey data collection and dissemination of federal survey statistics. The key objective is to develop alternative data collection frames and novel estimation strategies that inform federal statistics. The solution for this request should consider innovative approaches, such as machine learning and artificial intelligence (AI), new data collection methods, including but not limited to passive data collection and opportunities to utilize web scraping, and novel estimation strategies based on alternative data sources.

Background

This project is part of the National Secure Data Service (NSDS) Demonstration project. The NSDS Demonstration project under PL 117-167 calls for activities to "develop, refine, and test models to inform the full implementation of the Commission on Evidence-Based Policymaking recommendation for a governmentwide data linkage and access infrastructure for statistical activities conducted for statistical purposes, as defined in chapter 35 of title 44, United States Code."

The Advisory Committee on Data for Evidence Building: Year 2 Report recommended several focus areas for the NSDS to support an evidence-building ecosystem, including technical assistance leads to develop educational resources for data providers and data users (recommendation 3.5) and creation of tools and support to users in conducting secure, accurate, and scalable analyses (recommendation 5.5). This project will build on these recommendations.

Current State

Federal surveys are experiencing several challenges, including declining response rates, increased costs, competing collections, and decreasing resources to support survey operations. Agencies across the federal government are dealing with the challenges of how to maximize the available data resources and how to disseminate the results of these resources while protecting privacy and maintaining data utility. There are traditional approaches for federal survey data collection (e.g., sample-based surveys) that rely on sampling frames[1] that may be out of date, may become non-existent, or are currently limited to broader use due to statutory authorities. There are also a variety of other data sources that could be used to inform federal statistics.

The recent rule by the Office of Management and Budget, the Fundamental Responsibilities of Recognized Statistical Agencies and Units (RSAUs), requires RSAUs to carry out their four fundamental

---

[1] Sampling frames are used to identify the target population. Sampling from these frames enables researchers to select unbiased and generalizable results without collecting data from the entire population.

responsibilities and for supporting federal agencies to enable the RSAUs to carry out these responsibilities. For RSAUs to be able continue to carry out their responsibilities, new and flexible approaches for data collection, estimation, and dissemination need to be assessed.

Future State

Expanded use of readily available alternative data sources provides an opportunity to create new data collection frames for sampling when questionnaire data are needed or for use in the calculation of novel estimates when sampling is not needed or is not feasible. An approach can be developed where an NSDS would support federal agencies in the preparation and dissemination of trusted Federal statistics when traditional survey sampling methods may not be feasible. For example, a future NSDS may draw from administrative, sensor, stream, and/or web-scraped data to develop data collection frames. Such novel approaches would require assessment for fitness for purpose, data quality, and comparability to traditional methods as well as potentially new estimation strategies. Together, such novel approaches could enable the dissemination of federal statistics to continue if a traditional frame is unavailable due to a variety of reasons, including but not limited to a budget shortfall. Dissemination strategies for statistics derived from these new data collection frames will need to balance protecting privacy and maintaining utility. Determining how to communicate the uncertainty around statistics derived from these new approaches will also be critical to uphold data quality standards and transparency.

The Current Request for Solutions

This project seeks options for developing an open-source end-to-end alternative data collection frame, sampling strategy, novel estimation technique, and dissemination approach, using non-traditional data sources that could replace or augment traditional collection and estimation strategies for statistical data. The project should not be limited by statutory authorities, and should include detailed sampling methodologies, strategies for calculating novel estimates, and data dissemination tools. The project should consider the use of AI, API-translators, machine-learning, and other emerging technologies for collection, analysis, and dissemination of statistical data. The project should identify the risks associated with the proposed methods and provide mitigation strategies to handle these risks.

Respondents to this RFS should propose a solution that will be informative for the process of developing an alternative data collection frame, sampling strategy when questionnaire data are needed, novel estimation techniques when sampling is not needed or is not feasible, novel estimation strategies, and dissemination plans using non-traditional data sources. The project should include a case study that tests this data collection frame creation with the U.S. science and engineering workforce for key estimates such as following career pathways, science and engineering trends in education, and workforce credentials and training. This case study should demonstrate how an alternative data collection frame is created and how it is implemented for one population of interest.

Information Gaps

A successful project will identify:

- Alternative data sources that can be used to develop data collection frames and the use of the frames will be demonstrated with at least one test case on the U.S. science and engineering workforce;

- Strategies to utilize alternative data collection frames for sampling and calculating statistics from these frames that inform federal statistics with a flexibility needed in an evolving policy environment;
- Methods that can be applied to ensure that data meet data quality standards, including replicability for other policy relevant topics of interest.

Key Evidence-Building Considerations

Key focus questions to assess feasibility:

- How can using alternative data sources for data collection frame development inform and produce key federal statistics?
- What are limitations of using alternative data collection and analysis when used for statistical estimation and dissemination?
- How can novel methods be assessed to determine and convey appropriate-use information to users of the created data collection frames?

Deliverables

At a minimum, offerors will provide the following if selected for an award. Additional deliverables may be required.

- Monthly status reports on progress toward project objectives.
- Quarterly lessons learned that highlight what has been learned thus far, challenges with potential solutions, and any best practices that could be utilized to inform a future NSDS.
- Development and delivery of an open-source end-to-end approach that implements development of a data collection frame, sampling strategies, novel estimation techniques, analysis, and dissemination strategy for using alternative data sources to replace or augment traditional collection for statistical data, with a use-case that informs the U.S. science and engineering workforce estimates. This approach could use AI, API-translators, machine-learning, and other emerging technologies to create data collection frames.
- All source code and documentation for the project.
- A final report describing all non-traditional data sources and methods used to create the data collection frames and their limitations, a comparison to traditional data and methods, and a description of risks and potential mitigation strategies of the frames. In addition, the final report should describe sampling strategies when questionnaire data are needed and novel estimation techniques when sampling is not needed or is not feasible.
- A communications and outreach plan that is integrated into the services offered within an NSDS, with accompanying implementation plan, for communicating with stakeholder groups about the data collection frame development, sampling strategies, novel estimation techniques and dissemination strategies, messaging, and metrics to highlight the value proposition for using alternative data collection frames.