ATTACHMENT I – PROJECT TOPIC

Utilizing Privacy Preserving Record Linkage with Parent Agency Data and Statistical Agency Data to Inform Programs and Policies

Background

Throughout the federal government there are many initiatives to link data for evidence building. One way to support these initiatives is to build the infrastructure necessary for cross agency linkages, including developing data sharing agreements, conducting linkages while protecting privacy, and utilizing secure environments for analyzing the resulting linked datafile(s). To support these initiatives this project will serve as a research demonstration project for the National Secure Data Service (NSDS) to develop a data sharing agreement between a federal statistical agency and its parent agency, utilize a privacy protecting record linkage tool to link two disparate sources, and create an analytic dataset that can be used to answer questions that could not be answered with either source alone.

A data sharing agreement is necessary to provide structure and governance over how the data will be shared and how they will be analyzed once they are linked. This project will include the exploration of different data sharing agreement options to identify templates for future data sharing opportunities. In order to conduct the linkage, a privacy preserving record linkage (PPRL) tool will be selected. PPRL is a method that can be used to link de-identified data, using encrypted tokens and a trusted third party. PPRL will need to be performed in a secure environment that meets CIPSEA and FEDRAMP standards. Once the data are linked, the dataset will be available for analysis in a secure environment that meets the standards outlined above.

The two data sources to be linked using a PPRL tool are the National Center for Science and Engineering Statistics (NCSES) Survey of Earned Doctorates (SED) and the Principal Investigator (PI) data for awarded proposals maintained by National Science Foundation (NSF) in the NSF data systems. The SED is a census of all doctoral recipients in the U.S. The PI data contain personally identifiable information (PII) about principal and co-principal (PI/co-PI) investigators.

Once the data are linked using a PPRL tool and stripped of the encrypted tokens, they will be available for analysis in a secure environment. All output from analyses using the linked microdata will be assessed for disclosure risk before leaving the secure environment. The microdata could be used to understand the trajectory of U.S. doctoral recipients, which could inform programs and policies on education and funding. In addition, the linked data could answer questions about the receipt of an NSF grant as PI or co-PI after completing a doctorate. Then, using the linked data, one could assess if there are differences in receipt by various demographic variables, such as age at time of doctorate, race, ethnicity, state of residence, field of study, years needed to complete doctorate, and whether there were any changes by demographic variables and receipt of an NSF grant due to the pandemic. This work can also inform many of the objectives of the NSF learning agenda, such as:

- How could NSF leverage tools at its disposal— policies, strategies, programs, and so on—to increase the participation of these (most extremely underrepresented intersectional) groups in the STEM enterprise? Answers to these questions will help NSF identify best practices and align programs and policies toward closing gaps in participation in the STEM enterprise.
- How the work will facilitate analyses of data through a difference-in-differences approach (to measure differences in measures, such as proposals submitted by gender before and after the pandemic) and the specification of regression models as part of an interrupted time-series (ITS) design to determine changes that might be attributed to COVID— by modeling (and comparing) the expected pre-COVID and observed since-COVID trends, controlling for relevant factors.
- Findings will help NSF leadership and staff consider strategies for improving the efficacy and equity of the merit review process.

Objectives

The project objectives are:

1) Develop a data sharing agreement between a federal statistical agency and its parent agency.
2) Utilize privacy preserving record linkage methods to connect SED and PI data for awarded proposals in a secure server environment (e.g., a trusted third party managed by NCSES). This linkage would serve as an example of utilizing PPRL for data sharing and linking which will guide future interagency initiatives.
3) To demonstrate the ability to analyze linked data in a secure server environment. Once the linkage occurs the data will be analyzed to address key research questions and support the NSF learning agenda and other internal NSF stakeholder questions.

PPRL would be used to link the SED and PI data for awarded proposals. This work will highlight the ability to deploy a PPRL tool in a secure environment with a trusted third party, without the direct exchange of PII. The work must be completed by a CIPSEA designated agent and the secure server environment needs to meet CIPSEA and FEDRAMP standards. This work will inform linkages across the federal government, using the development of agreements and deployment of PPRL as a model to improve the availability, quality, accessibility, and interoperability of data sharing. This work will demonstrate the technical infrastructure that would be needed to allow linkages to occur quickly and efficiently for specific projects across federal agencies. In addition, it will demonstrate the ability to analyze the linked data on a secure platform to answer key research questions.

Information Gaps

This project will identify key components necessary to inform a future/potential NSDS including:

- The steps needed to develop a data sharing agreement between a federal statistical agency and its parent agency.
- What is needed to conduct PPRL to link these sources without ever exchanging direct PII.
- What is needed to build an infrastructure to inform linkages across the federal government, using the development of agreements, utilization secure server environment, and deployment of PPRL as a model to improve the availability, quality, accessibility, and interoperability of data sharing and linking.
- What is needed to assess the feasibility of analyzing linked data in a secure environment to support evidence-based policymaking.

Key Evidence Building Considerations

- Key focus questions to assess innovation in the following areas: data acquisition, data security, data linking, privacy, and engagement:
    - What are key challenges with locating, acquiring, accessing, linking, and using disparate data and information?
    - What are the lessons learned to using the presumption of accessibility in the Evidence Act?
    - Which novel techniques for data, privacy, and confidentiality protections can be used?
    - What linked data models and estimation methods are best suited to fill information gaps?
    - Are the resulting linked data and models suitable to inform policy discussions and to make data available more equitability?
    - What types of collaboration and stakeholder engagement are needed to help inform these questions, data, and analysis?
    - What mechanisms are needed to access the resulting data that uphold privacy requirements?

Deliverables

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

- Develop a data sharing agreement between a federal statistical agency and its parent agency, including lessons learned to inform future efforts.
- Selection of a PPRL tool that meets certain minimum requirements, including but not limited to successful past performances with documented high precision and recall. The tool should demonstrate performance linking with identifiers such as name, email, and address.
- Creation of two submission files that have been standardized for linkage
- All code (clearly documented), linked data file, documentation of linkage process, documentation of linked data file (including codebook), and any other data/documentation created under this award.
- Report summarizing findings from statistical analysis using the linked SED-PI data
- Final report summarizing all stages of the project and offering lessons learned for a potential future NSDS.