ATTACHMENT I – PROJECT TOPIC

Artificial Intelligence for Enhancing Data Quality, Standardization, and Integration for Federal Statistics

Key Objective

In 2016 the concept of a National Secure Data Service (NSDS) was introduced by the Commission on Evidence-Based Policymaking for the purposes of providing a platform for shared government services to streamline data access, data linkage, statistical analysis, and privacy protections. Pursuant to the 2022 CHIPS and Science Act and further recommended by the Advisory Committee on Data for Evidence-Building (ACDEB), a National Secure Data Service Demonstration (NSDS-D) is required to facilitate the development of statistical and data infrastructure and to inform efforts at establishing a full NSDS. The objective of this project is to explore the use of artificial intelligence (AI) to assist in the processing, formatting, standardization, and integration of data to support activities that are integral to statistical agencies' production of high-quality statistics and data products. These efforts will be integrated to develop tools and services to support a future NSDS that supports several federal agency priorities, including capacity building, data access and sharing, and privacy protections.

Federal agencies often use data from disparate sources lacking common formatting and standardization. Increasingly, this includes nontraditional data. While nontraditional data may take many forms, it typically lacks the strong structure of traditional survey data, making it challenging to use. There are also privacy concerns associated with nontraditional data. This project will explore the development of a set of data processing tools, utilizing AI, to enhance data standardization and integration activities that are central to the data quality requirements for the federal statistical system and beyond. Recent AI innovations can benefit data preparation for statistical analysis and can be a part of this data service's toolkit offering. The envisioned future shared service would streamline data processing, promote data standardization, and provide data quality guidance.

Background

Many datasets face issues with consistency, integrity, and usability. Such challenges are common even with classical, structured survey data, and preprocessing data for quality control is often a prerequisite for statistical analysis. Nonetheless, institutions are increasingly augmenting traditional, survey-based methods of data collection with newer, nontraditional methods of data collection. Examples of nontraditional data include cellphone data, satellite and global positioning system (GPS) data, radio frequency identification (RFID) data, administrative records, and transaction records. In the era of electronics and computing, the movements and activities of people create digital footprints. This has created an abundance of potential nontraditional and often unstructured (i.e., not in classical tabular format) data. Nontraditional data, unlike traditional survey data, have not been designed with statistical end use in mind but rather as a byproduct of other activities. Hence nontraditional data typically face even more challenges in processing, standardization, and privacy protections.

A data standardization and integration service aligns with the aims listed in several federal data quality mandates and guidelines. The Office of Management and Budget (OMB) has outlined several directives

and standards detailing minimum requirements for federal statistical agencies. Following the Information Quality Act, agencies are tasked with maximizing the "quality, objectivity, utility, and integrity" of information, particularly statistical information. The National Academies' Committee on National Statistics (CNSTAT) outlines five principles and ten practices for guiding federal statistical activities. Several of these principles and practices address aspects of data quality, integrity, and credibility. The practices also call for coordination and collaboration across statistical agencies as well as openness about data sources and data limitations. A well-designed data standardization and integration service enhances the capacity of statistical agencies to collaborate on statistical projects and to provide high-quality data and analysis in adherence to these requirements and guidelines.

Tools and methodologies exist for addressing various data challenges. A data standardization and integration service would provide federal statistical agencies with a central hub to access these tools to perform these tasks. A well-designed data service would accelerate data processing and formatting, standardizing data, and facilitate collaboration between statistical and nonstatistical federal agencies. Ideally the service would provide mechanisms for addressing the three domains and eleven dimensions of data quality outlined by the Federal Committee on Statistical Methodology's (FCSM) "A Framework for Data Quality."

Recent advances in AI and machine learning offer powerful tools for handling data and deriving insights, especially from unstructured data sources. Developing AI tools that address federal and public needs is an active effort, with the 2020 National AI Initiative Act outlining the creation of the National Artificial Intelligence Research Resource (NAIRR) task force to oversee this endeavor. Part of NAIRR's directive is the establishment of AI tools that further the capabilities of the workforce. Here, effective AI tools that enhance data quality, standardization, and integration would expand the capacity for statistical activities. The National Academies has hosted several discussions covering the use of AI for optimal data collection, data quality control, and real-time data monitoring. AI data processing has the potential to offer, among other things, text and natural language processing support, automated data formatting, and anomaly and quality control detection. A data processing toolkit that utilizes AI can be an integral component of a data standardization and integration service. In general, a strong data standardization and integration framework will enable the pooling of understandings and lessons learned from different agencies. This will ultimately standardize user experiences, improve analyst productivity, and reduce learning curves, enabling more rapid and meaningful statistical analyses. This will explore 3-5 publicly available data types that may be expanded to explore other types of data in the future.

Information Gaps

Information gaps to be identified include:

- Summary of common data quality issues with traditional and nontraditional data faced by federal statistical agencies
- Description of types of traditional and nontraditional publicly available data as well as standardization procedures for the different types
- Frequently encountered data linkage and integration challenges by types of data
- Lessons and best practices that individual federal statistical agencies have learned and how might these be developed into a set of guidelines for all agencies
- AI technologies that are currently available and can address these data quality needs
- The hardware/infrastructure needs for developing the service

Key Evidence Building Considerations

The following questions may guide the development of a service addressing data quality, standardization, and integrity:

- What data standardization issues exist and what can be addressed with AI in a toolkit?
- How can AI tools addressing data quality challenges be packaged into a service that is user-friendly, accessible, and amenable to regular updates?
- How can efforts at building a data standardization and integration service be coordinated with stakeholders?

Objectives

The objectives of this project are to:

1. Develop a framework for data standardization and integration services, including but not limited to the following topics. Additionally, outreach should be conducted with statistical agencies that have already considered these issues.
    - Identify common traditional and nontraditional data types (e.g., survey data, administrative data, cellphone data, etc.)
    - Identify 3-5 data types (to be determined) for standardization and integration needs (e.g., string formatting for text data, linking survey and nontraditional data, etc.)
    - Identify ethical and privacy concerns and risks, particularly with respect to utilizing AI to develop tools and services for data standardization and integration.
    - Identify AI applications that address data processing, standardization, integration, and privacy concerns
    - Identify computational and software tools for service deployment (e.g., python, R, etc.)
2. Leverage or design tools and services (open source preferred) to facilitate data processing, standardization, integration, and privacy protection that can ingest traditional and nontraditional publicly available data types. Elements of the toolkit may include, but are not limited to:
    - Scraping and conversion of data from noneditable to editable formats (e.g., PDF or images to spreadsheet or tabular data)
    - Web data scraping tools to facilitate and automate data collection from websites and APIs
    - String formatting and natural language processing tools (e.g., topic clustering, sentiment analysis, etc.)
    - Data event detection such as outlier, anomaly, or change point detection
    - Traditional and nontraditional dataset merger assistance
3. Develop documentation, use guidelines, and best practices:
    - Provide a checklist of best practices when processing and standardizing data
    - Provide users a summary of tools available and scope of applications
    - Document performance measures of AI tools (e.g., speed, computational resource usage, data size constraints, etc.) for comparison with standard approaches
    - Provide users with guidance and direction on addressing privacy concerns
    - Inform users about the applicability of data, the data context, and providing caveats about limitations with datasets, based on the data quality domains and dimensions outlined in the FCSM Framework for Data Quality

4. Package tools and services into an accessible service and user-friendly interface that may be included in a future NSDS but also accessed across different agencies and data platforms.

Deliverables

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

- A framework plan that is informed by current activities and best practices across statistical agencies for the development of a data standardization and integration service and toolkit that leverages new or existing activities, tools, and processes that align with the FCSM Data Quality Framework.
- A set of user-friendly tools (i.e. toolkit) packaged into a service addressing or enhancing data processing, standardization, integration, and privacy protection.
- A user interface that is intuitive and readily integrated into a future NSDS and accessible to different agencies.
- Documentation that includes but is not limited to a set of guidelines, best practices, and caveats about using the toolkits and service.
- Evaluation deliverables
    - Monthly progress reports
    - Regular meetings with federal staff
    - Quarterly lessons learned
    - Final report detailing project goals and final outcomes including but not limited to tradeoffs in data quality and risk mitigation strategies