ATTACHMENT I – PROJECT TOPIC

Data Access Alternatives: Artificial Intelligence Supported Interfaces

Key Objective

The key objective of this effort is to create and test machine-learning-backed or "artificial intelligence" (AI)-backed user experiences with federal statistical data. This experience shall improve on the current state of user interactions based on obtaining answers to questions via search engines or emailing federal staff or contractors. This project seeks to develop and pilot an AI chat bot (or the like) that answers users text queries submitted via an interface. Answers should be obtained from public statistical data of federal statistical agencies in this project.

Background

The CHIPS and Science Act, PL 117-162, was signed into law in August 2022. Section 10375 establishes a National Secure Data Service Demonstration Project (NSDS-D) 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 recommends several approaches for the NSDS that focus on supporting a high-quality user experience. Recommendation 3.3 states that "The NSDS should identify opportunities for automation of its "intake process," providing a high-quality user experience while focusing staff effort on complex user needs". Recommendation 3.4 states "The NSDS should employ data concierges to help users refine their research projects, discover relevant data, and acquire access to that data."

Current state

Statistical agencies produce a large amount of data and analysis every year. These data can answer questions such as "How do most people die?" and "How many employed engineers with doctorates live in Arkansas?". However, simply typing these questions into an internet-wide search engine or search bar on a statistical agency website will not always return the direct information requested. In addition, it may be difficult to know with an internet-wide search if the result is based on an statistic from federally published data or other non-federal data. If the user happens to use the terms the agency used, then relevant pages may be returned. As long as a user knows the exact terminology, search engines excel at returning relevant results. However, even then a user may need to then read the page looking for the desired specific statistic, which may be buried in a table. However, many times, users do not know the correct terminology or how to structure the requested search to obtain the statistics. For example, for the first question, a user would need to have typed in "leading causes of death" to get the relevant page on the National Center for Health Statistics webpage.

Future state

The expansion of machine learning, including natural language processing and AI may provide an updated approach to directly answering user queries related to public statistical information. One option one can imagine is instead of having to know exactly what to search or where to look, users could ask a chat-bot like entity their question – for example, how many employed engineers with doctorates live in Arkansas?

This project seeks to build a pilot tool to respond to users' queries using publicly available data on statistical websites. The tool should be designed to be included on a public-facing federal (.gov) webpage. The tool development should employ best practices related to accuracy and reproducibility and produce metrics that allow for transparency.

At the minimum the tool should achieve the following:

- Natural language processing of user queries.
- Return both text and graphics, depending on the query.
- High accuracy in any returned statistics, with "I don't know" being returned for questions that do not reach this metric. (Responders should define "high" for their proposed solution.)
- Source citation and link with each statistic returned.
- Potential seamless integration with existing federal websites.
- Transparency about the limitations of the model results.
- Consistency with all federal policies governing agency use of AI.

This project shall also identify the risks associated with the tool and suggest mitigation approaches. The pilot should identify how portable and scalable the resulting tool is, along with technical requirements for both.

Respondents should propose a solution that will use public federal statistical data from a limited number of statistical agencies, including the National Center for Science and Engineering Statistics and others to be determined[1], to answer a series of query questions requesting existing statistics or graphics. Proposers are encouraged to propose other future states that will meet the key objective of increasing user access to federal statistical information through a machine-backed interface.

Information Gaps

This project will identify:

- If AI or other machine-learning approaches can be utilized to improve user access to federal statistical information.
- How needed services and support could potentially be integrated into a data concierge service for a National Secure Data Service.
- The extent to which information returned by a chatbot maintains fidelity to the requirements of the Information Quality Act.

Key Evidence Building Considerations

[1] This pilot will focus on small agencies (<500 full-time permanent staff).

- Key focus questions (address one or more) to assess innovation in the following areas - user engagement and customer service:
    - How can a machine-learning-based chat bot increase user access to public federal statistical data?
    - What role can this approach play in a tiered access system?
    - How would this tool assist in a data concierge service?

Deliverables

At a minimum, offerors will provide the following if selected for an award. Proposers should outline the additional deliverables they will provide in the provision of this solution.

- Project plan outlines key milestones and timelines.
- Monthly status reports on progress towards project objectives.
- Quarterly lessons learned based on what has been learned during the last quarter.
- Final report detailing tool capabilities and limitations, data sources used to train models and derive responses to queries, and a compilation of lessons learned.
- The tool with technical documentation required for maintenance and replicability. The tool should be developed using or resulting in open-source software or code, which should be made publicly accessible (say on github or equivalent) and accompanied by complete documentation of the tool.
- A set of recommendations to enable AI-driven search and discovery of content on federal statistical agencies' websites.
- A transition plan to transfer the tool to an .gov platform.