ATTACHMENT I – PROJECT TOPIC

AI-Ready Data Products to Facilitate Discovery and Use

Key Objective

More and more Americans are using readily available generative artificial intelligence (AI) technologies to guide their everyday decisions. It is critical that the responses from these models reflect accurate, unbiased information. As the nation's trusted sources of information on a wide array of topics—the economy, transportation, health, education, and more—federal statistical agencies and units should explore ways to make their statistics more discoverable and ingestible by generative AI.

The objective of this project is to explore how a future National Secure Data Service (NSDS) could provide shared information and tools for making statistical data products more readily ingestible by AI technologies. These resources could support federal agencies and their partners throughout the data and evidence ecosystem as they balance the risks and rewards of leveraging generative AI to expand the reach of their statistics.

An important first step is figuring out how AI tools capture and report federal data. From there, agencies could apply best practices and test solutions for transforming their data assets from machine *readable* formats, as required by the Foundations for Evidence Based Policymaking Act of 2018 (P.L. 115-345) (or the "Evidence Act"), to machine *understandable* products, as necessitated by AI algorithms. The results of this project could be folded into a more comprehensive toolkit, covering AI applications for data collection, processing, analysis, and dissemination. Finally, this effort could offer critical touchpoints for tackling the challenges and opportunities of other emerging technologies.

Background

What is AI-ready data?

Title II of the Evidence Act (or the "OPEN Government Data Act") "requires public government data assets to be published as machine-readable data." That is, "data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost." To be leveraged by generative AI technologies government data assets should not only be machine readable but machine understandable.

The Department of Commerce recently stood up the AI and Open Government Data Assets Working Group to "[develop] guidelines for publishing Commerce data that can be consumed by emerging AI technologies." The Department describes machine-understandable, "AI-ready" data like this: "data that [are] enriched with contextual metadata and organized in interpretable standard formats." By making its data AI ready, "AI models can then better interpret Commerce data, link them to similar data, and return accurate results from authoritative sources." This posture of AI readiness applies to data assets held by other federal agencies and by partners throughout the data and evidence ecosystem.

How does this connect to the NSDS?

The CHIPS and Science Act, Section 13075(c), establishes an NSDS 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 [the Evidence Act]." Furthermore, this work is to "be established in consultation with the Office of Management and Budget and the National Artificial Intelligence Initiative Act of 2020 Interagency Committee." These connections, along with recent AI Executive Orders and guidance, highlight the possibilities and potential pitfalls of AI.[1]

The NSDS-D is launching projects to investigate AI opportunities. For example, the Request for Solutions on "Data Access Alternatives: Artificial Intelligence Supported Interfaces" "seeks to develop and pilot an AI chat bot (or the like) that answers queries submitted via an interface. Answers should be obtained from public statistical data of federal statistical agencies in this project." This project on AI-ready data products provides an important building block for ensuring the success of such an interface.

Project Overview

The project involves two main parts:

1. **Assessment of the accuracy and timeliness of federal statistics in generative AI tools.**
    a. **Landscape analysis.** Perform a landscape analysis of the machine "understandability" of federal statistical agencies' public data products. This should reflect a cross-section of statistical agencies (with a minimum of six agencies including those involved in the case study and developing the AI readiness tool described below), different information types (e.g., website, XLS, PDF, graphics, interactive tabling tool, API), and several of the most common generative AI engines.

    Results should include (1) an overview of the accuracy and timeliness[2] of how different AI models capture and report publicly available federal statistical data, (2) an inventory of current industry best practices for optimizing data representation in generative AI tools, and (3) a repeatable assessment of how well agencies' data products align with those practices, including a method for ranking data products from most AI ready to least AI ready.

    b. **Case study—Bureau of Economic Analysis (BEA).** The case study should include (1) a prioritized list of BEA data products for AI readiness transformation (developed in coordination with BEA and informed by the AI readiness assessment described above) and (2) for those data products selected as "high priority," specific recommendations or alterations that would maintain the human-readable quality of the products and enhance their ability to be consumed by machines.

2. **AI readiness tool with prototyping and replicability testing.**
    a. **Tool and prototype.** Develop a technical solution for transforming federal statistical products into machine understandable, AI-ready data. This solution should implement industry best practices for data management and present agencies' data and metadata to

[1] For more information, see Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" and "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (M-24-10).

[2] "Timeliness" in this context is the frequency at which AI models review and present data that has been updated. If this is not regular enough, the results provided may not reflect the latest available statistics.

optimize their consumption by generative AI. The utility should take a generalized approach that could be applied to the data products of any federal statistical agency. This may be a stand-alone solution for optimizing machine "understandability" or may yield information that could be consumed by both humans and machines.

This prototype tool would initially be applied to high-priority BEA data products, as determined during the case study, and would be tested using protocols established during the landscape analysis. These tests should validate that the solution improves the accuracy and timeliness of how generative AI tools represent the data from the chosen products. The resulting AI-ready data must meet BEA requirements to be hosted on the agency's website.

b. **Testing and replicability.** Apply the prototype tool to data products from another statistical agency (or agencies), run tests for machine understandability (including semantics), document differences from the BEA prototype, update the tool (as appropriate), and recommend next steps for wider use.

Information Gaps

Key outcomes of this project include the following:

- Support federal agencies in building capacity to (1) evaluate generative AI technologies' access to and use of their statistics, (2) identify high-value assets to make AI ready, and (3) provide solutions for doing so.
- Develop a tool that can be used by federal agencies to transform data assets into machine-understandable, AI ready data products.
- Lay groundwork for (1) expanding decisionmakers' access to federal data through widely available generative AI tools and (2) connecting to an AI interface tailored to respond to user prompts with authoritative statistical information, potentially as part of a future NSDS.

Key Evidence Building Considerations

Key questions include the following:

- How is generative AI currently using trusted federal datasets to respond to user prompts?
- How can federal agencies improve their data dissemination methods and ensure the integrity, accuracy, and timeliness of their data when used in AI applications?
- How can agencies identify which data products would be most impactful if made more ingestible by AI algorithms?

Deliverables

As AI technologies rapidly evolve, it is essential that this project keeps pace with advances in the field and aligns with best practices and standards as they develop. In addition, project offerors should demonstrate AI technical expertise as well as in-depth knowledge of federal statistical products.

At a minimum, offerors will provide the deliverables described below if selected for an award. For more information, see "Project Overview." Additional deliverables may be required.

- Biweekly or monthly status meetings with the project team.
- Monthly updates on progress toward project objectives, including quarterly highlights of lessons learned.
- Interim reports describing (1) the landscape analysis and case study and (2) the BEA prototype and replicability testing as well as (3) a final report that can be used as a roadmap for other federal

agencies to increase AI readiness, highlighting successes, challenges, and lessons learned during this project.

- An AI readiness assessment that could be a shared resource for any agency looking to test the machine understandability of its public data products.
- A generalizable and sustainable technical solution that transforms statistical data products into AI-ready data and step-by-step documentation that describes how to apply the solution, using examples from the BEA and replicability tests.
- Recommendations on hosting and sharing reports, tools, and documentation developed in this project to encourage re-use and refinement as part of a shared service within a future NSDS.
- Communications plan to promote the AI readiness solution and resources with federal agencies and their data partners.