## ATTACHMENT I – PROJECT TOPIC

# Creation of Synthetic Data for the Survey of Earned Doctorates and Development and Use of Verification Metrics

### Key Objective

The objective of this project is to produce a synthetic data file for public use to support a tiered access model, explore the use of synthetic data for evidence-building, and test the use of verification metrics in validating estimates produced from synthetic data. A synthetic dataset does not contain the exact records of the original dataset, but instead retains the statistical properties of the original dataset, preserving information useful to researchers and their queries. The anonymity of the original dataset is not compromised since synthetic records do not correspond with real ones.

The Survey of Earned Doctorates (SED) is an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year. The SED is sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF) and by three other federal agencies: the National Institutes of Health, Department of Education, and National Endowment for the Humanities. The SED collects information on the doctoral recipient's educational history, demographic characteristics, and postgraduation plans. Results are used to assess characteristics of the doctoral population and trends in doctoral education and degrees. To address disclosure concerns, a synthetic datafile option is being explored for public use. This project would build on this work to inform the National Secure Data Service Demonstration Project (NSDS-D) and support tiered access through the development of a publicly available dataset that could be used without the barriers involved in accessing the restricted-use data.

### Background

America's DataHub Consortium brings together capabilities and infrastructure to securely fill information gaps and to take on key analytic questions and evidence building challenges. As demand for access to confidential federal data assets increases alongside novel analytical approaches, privacy protections must be in place to ensure the protection of privacy and the confidentiality of the data. Use of synthetic data can reduce disclosure risk while allowing data users to access microdata for research and other statistical purposes. The production and use of synthetic data for the SED is being explored to increase the utility of the data while ensuring strong privacy protections.

The SED is a census of all individuals receiving a research doctorate from U.S. academic institutions in a 12-month period from July 1st to June 30th of the data collection year. The SED is conducted annually and collects demographic information of the respondent; educational history and degree obtained; postgraduation plans; and financial aid. These data are highly valuable in studying educational trends and educational differences by demographic groups including studies of equity. Currently, a restricted-use version is available for researchers who are approved for use, who enter into a licensing agreement with NCSES, take annual training, and use the data in a secure virtual environment. In addition, there is a publicly available data tool that allows researchers to submit queries to produce tabulations and a public use data set with limited number of variables.

The proposed availability of a synthetic dataset would provide an additional tiered access option that would allow researchers to use SED microdata without an application or restricted use data license. This option could prove valuable to researchers in making maximum use of these data while enabling the government to ensure privacy. The production of synthetic data, though, can prove challenging and resource intensive.

The following steps will be completed to explore the development of synthetic data:

1. Initial assessment to determine what variables for a synthetic dataset would be most useful for current and potential policymakers, researchers, data users and other stakeholders. This assessment would involve outreach to these policymakers, researchers, data users and other stakeholders.
2. Open-source packages, such as R packages, would be utilized to create a synthetic version of the SED to produce a public-use microdata file. All contract staff accessing data will need to become CIPSEA agents and take annual data security training.
3. Once produced, the dataset would be evaluated to assess quality and disclosure risk. Use cases for this synthetic data would be identified to assist in determining how the synthetic data could be optimally utilized by policymakers, researchers, data users and other stakeholders.
4. To assess the alignment of the synthetic data with the restricted data and to assess the quality of estimates produced from the synthetic data, verification metrics would be developed so that policymakers, researchers, data users and other stakeholders would be able to request them on an as needed basis. Note: no estimates based on the restricted data would be shared but rather a metric indicating alignment of the estimates (e.g., confidence interval overlap, a flag indicating whether the synthetic estimate fell within the confidence interval of the true data).
5. A plan for dissemination of the synthetic data, verification metrics, and public messaging regarding this new public-use data product would be developed in conjunction with stakeholder outreach and feedback. Specific attention in the dissemination and public messaging should be focused on the role of the synthetic data in the SED's tiered access model and how to describe synthetic data to users.

## Information Gaps

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

- If a synthetic dataset can be produced on a demographic dataset using open-source packages.
- How verification metrics can be used to inform the quality of synthetic data in evidence-building.
- A path to produce verification metrics for an agency's restricted data and transfer those metrics to a centralized shared server. Use cases for a synthetic version of a high-value demographic dataset.
- Model for dissemination of a synthetic dataset in a tiered access model and messaging surrounding potential uses.

## Key Evidence Building Considerations

- Key focus questions (address one or more) to assess innovation in the following areas: data acquisition, data security, data linking, privacy, and engagement:
  - Which novel techniques for data, privacy, and confidentiality protections can be used?
  - Are the resulting data and models fit 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.

- Monthly report, using NCSES template, to document progress.
- Stakeholder outreach summary.
- All code (clearly documented), documentation of synthetic data methodology, documentation of data quality assessment, and any other data/documentation created under this award.
- Documentation of verification metrics and alignment with true estimates.
- A report describing the lessons learned through this project including but not limited to the creation of synthetic data and whether the resulting data and models fit are able to inform policy discussions.  In addition, the report should describe how this approach could inform a tiered access model and contribute to a potential National Secure Data Service.