The Standard Application Process is an important step forward for improving the use of very valuable data collected by the Federal government while strengthening protection of those data by making applications processes more transparent and making available information about analysts using them. The Federal Register Notice (FRN) 2022-00626 requested comment on "The Interagency Council on Statistical Policy's [ICSP] Recommendation for a Standard Application Process (SAP) for Requesting Access to Certain Confidential Data Assets" with specific questions about such topics about metadata standards.

The National Center for Education Statistics (NCES) reviewed the materials and has the following comments.

The first is about the scope of the standard application process (SAP). Page 4 of the ICSP report indicates that statistical agencies or units designated under CIPSEA would use SAP. Related to this, it is our understanding that data collected under other confidentiality pledges by agencies or units would also be included. Is that correct and would data collected under other types of pledges – including those with no pledges of confidentiality - be required to be included in SAP?

On the same page, the ICSP report suggests that the statistical agency or unit who collected data would be responsible for all access to and use of those data. Is that correct? If so, can the second paragraph on page 4 be clarified to indicate explicitly that statistical agencies or units control confidential data that they collected throughout the entire SAP process and that the statistical agencies are the one who provide such data to requestors that come to them through the SAP portals. We believe this is the overarching intent of the Evidence Act.

Also on page 4 of the ICSP report and the FRN, there is language to suggest that the statistical agency that collected data or developed a derivative data product accessed through the SAP would retain sole responsibility for storing, sharing, protecting, and curating these data throughout the full SAP process. The SAP is meant to improve discoverability of data, provide a consistent and predictable process for applying for data, and an index of who accessed the data and how they were used. Is this a correct interpretation of the material? 44 USC § 3520(d)(3) seems to make this clear but direct statements in the SAP establishment process and clarifying language in regulations in development for the Evidence Act about statistical agency control of data they collect or produce would help make the point more clear.

Plans are described in the ICSP report for establishment of a Governance Body and a Project Management Office (PMO). Are there plans to request from Congress resources to support the Governance Body? Given the volume of data that will be catalogued for the SAP process, our expectation is that there will be a large number of challenging issues to manage for the overall Federal Statistical System (FSS). The issues will be expanded because many applications will generate novel issues related to comingled data. Without at least some dedicated resources, the Body may not work efficiently which will be important given clearance schedules being proposed.

The PMO is meant to manage the SAP interfaces, curate a federal data inventory and related metadata, ensure that SAP required transparency features like lists of applicants and applications have proper Privacy Impact Assessments (PIA) and System of Record Notices (SORN), and manage security clearance processes. We recommend that dedicated staff and funding resources be established to operate the PMO and that those resources be made transparent to the public. We expect that these operations will represent significant investments by the Federal Government.

Page 7 discusses different roles in the SAP process. It might be helpful to clarify any role an agency OCDO or the OCDO Council might have at this point. 44 USC 3520(d)(3) delegates decisions about statistical data and data products to the agency statistical officials. This is often the head of a statistical agency or unit. However, consistent interpretations about OCDO roles in materials related to statistical data sharing would be useful across Evidence Act related materials.

Given the role of the PMO, would it be able to review existing SORNs under which data are stored by statistical agencies or units to determine if government-wide SORNs would be useful? SORNs describe both how data will be stored and how they can be accessed and used. Many statistical collections are undertaken for consideration of a particular research question or program evaluation like understanding infant health. Different SORNs across the government limit access to the statistical data for research projects that are specific to what the data were collected for originally like infant health. The limitation can make sense in terms of privacy concerns and agency resources needed to manage external access to the data. However, it can also limit important research. Staying with the infant health example, if an agency's SORN indicates that the data can only be used to study ways to improve infant health, then those data can be prohibited from being shared with analysts trying to understand how infant health might affect kindergarten academic performance. After a centralized review by the PMO and discussions with agencies, a government-wide SORN clarifying that statistical data can be used for research for which the data were not originally collected could help address a currently heterogeneous approach to interpreting access determinations. If such an approach is not feasible, the metadata in the SAP inventory should have a category indicating when data can only be used for specific purposes.

As part of the inventory process, there is discussion in the ICSP document about statistical agencies constructing new data assets. We would ask that clarification be provided about whether this would be a requirement or a recommendation. Constructing new data products would require additional resources for many statistical agencies and we do not think this should be a feature that we promise applicants. Such work could also take longer than the SAP review times envisioned in the document.

Related to timelines, pages 24 and 25 of the ICSP report provides clearance timeline expectations. How were the deadlines proposed in the situations described points a. – g. determined? Some agencies might have resource constraints that make it difficult to meet the deadlines described in some of these situations. Maybe these deadlines should be extended to accommodate agencies with more severe resource constraints – especially the 3-week deadline listed in f.?

Related to the application process, data access has a 4 tiered approach in the SAP proposal. This makes sense. The ICSP report notes that for access to some data assets, security clearances are needed. We will need to take clearance requirements into account when assigning lead agencies in multi-agency applications. Many agencies do not have processes in place for security clearances for data access. Also, should the SAP process include expected completion times for tier 1 and tier 2 security clearances?

Finally, comments were requested for metadata standards to facilitate discovery by researchers. This is important for SAP to work as expected. The FRN asks for ideas about possible standards to adapt. Some points to consider by OMB and ICSP on metadata.

1. The Committee on National Statistics (CNSTAT) in the National Academies of Science, Engineering, and Medicine (NASEM) recently produced a report for the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES). The report makes extensive recommendations for metadata including use of Data Documentation Initiative

Alliance (DDI) standards. Please see "Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies (2021)" at - https://www.nap.edu/catalog/26360/transparency-in-statistical-information-for-the-national-center-for-science-and-engineering-statistics-and-all-federal-statistical-agencies.

2. Could consider building from metadata standards developed or under development for Data.gov managed by GSA.

3. Metadata can be extensive, but we think what would be ingested into the SAP catalogue would include at least the following key pieces of information:

a. A sentence or two about why the data were collected

b. Whether data are publicly available without clearances or need a given tier of clearance

- If data are publicly available without clearances, links to those data if possible.

c. Mode of data collection used to capture the data

d. Who the reference population is for the data – who the data are meant to represent

e. Who provided the data in terms of respondent types

f. If the data are based on a census collection or on samples.

- If the data are sample based, are weights needed?
- If the data are sample based, was the sample a simple random sample or will variances need to be adjusted to account for complex sampling?

g. Number of cases available in the data for analysis (may require some rounding for confidentiality reasons).

h. If the data are associated with Personal Identification Keys (PIKs) with particular data in other agencies