

# COPAFS hosted Tiered Access Workshops

Funded by the Alfred P. Sloan  
Foundation

Cynthia Clark  
COPAFS



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Proposal for Workshop on Tiered Data Access

- Motivated by Foundations for Evidence-based Policy Act
- Goal to facilitate OMB Statistical Policy role to develop directives to enable legislative requirements
- OMB could benefit from insights from constituents interested in research access to statistical and program data sets
- Submitted proposal to Sloan Foundation to convene workshop
- Activity in line with COPAFS objective to encourage discussion among professionals on issues in federal statistics



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Invited Planning Meeting - September 2019

- Provide insights in development of FRN on data access – ideally specifying different levels of access
- Convened attendees from federal statistical agencies, contractors, and statistical data users
- Provided background on the Evidence-based Policy Commission's recommendations for data access & the Foundation Act to attendees
- Used NYU data lab as an example of a structure for data access
- Panel of government leaders discussed data access challenges



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Recommendations from discussion of key Q's

- OMB should encourage experimentation among different options for access
- Identify & address legislative barriers to use of data for program & policy evaluation meeting CIPSEA definition of statistical purpose
- Develop a continuum for mitigating risk in the 5 Safe constraints
  - Setting
  - People
  - Projects
  - Output
  - Data



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Recommendations from discussion of key Q's (continued)

- Conduct research into possible approaches for access in lieu of public-use microdata files
- Research approaches to quantification and estimation of risk of disclosure of identifying information
- Conduct holistic review of individual agency products to balance data utility with confidentiality and privacy concerns



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Invited Workshop – January 2020

- Designed around different frameworks for data access
- Focused on Five Safes, but references others
- Explore each component in detail & then review combinations of the five categories at different levels
- Invited participants included four attendees from the September planning workshop and three participants who had implemented aspects of the Five Safes (two international)



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Workshop Agenda

- Presentation on the Five Safes and their use in the NYU data lab
- Small group discussion of four of the safes – setting, people, project & output – with goal of developing a continuum of options across each category
- Review of Statistics Canada Confidentiality Classification Tool that provides quantitative assessment of the sensitivity of data
- Examination of three-tiered UK Data Service model
- Discussion of topics desired in any access protocol



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Safe Setting

- IT based solutions using encryption can facilitate a safe setting
- Continuous monitoring needs to be employed
- Physical requirements should be the exception but not the rule



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Safe People

- Training protocol is necessary & should have an ethics focus
- Affiliation with a recognized research or public institution is a means of ensuring safe people
- Individual and institutional sanctions that involve reputational cost are critical, but need to employ due process for violations
- Different levels of data sensitivity should be reflected in both training and sanctions



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Safe Projects

- Employ a sound research design involving ethical considerations and IRBs as appropriate
- Need to assess number and results of data linkages
- Reflect lawful/constitutional uses
- Useful to develop repositories of research projects for specific data sets to facilitate research development



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Safe Outputs

- Assess
  - likelihood of re-identification
  - likelihood of malevolent intruder using specified data source
  - Degree of harm/sensitivity
- Continuum of safe outputs
  - Public data – no output review
  - Researcher controlled output
  - Secure facility with restricted and reviewed output



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Stat Canada Confidentiality Classification Tool

- Quantitative assessment of the sensitivity of data
- Combines probability of different types of disclosure
  - Attribute
  - Identity
  - Inferential
  - Residual
- And Severity of harm
  - Severe, high, medium, low, negligible
- In matrix scoring level of confidentiality – 1 to 9 (highest risk)



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Stat Canada Confidentiality Classification Tool (continued)

- Use the CCT score with the accreditation level (safe people) to determine the safe setting
- Safe Setting approaches for research used by Stat Canada
  - RDCs
  - Remote system access to make requests
  - Synthetic data sets
  - Public Use Microdata Files (PUMFs)



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# UK Data Service Three-Tiered Model

- Open
- Safeguarded – researcher assert that project has a research purpose – similar to a limited public use microdata file
- Controlled –different types of access utilizing combinations of levels of Five Safe criteria
  - Safe data – treat data to protect confidentiality
  - Safe people – educate researchers to use data safely
  - Safe projects – research projects for public good
  - Safe setting – secure lab environment for personal data
  - Safe outputs – secure lab projects



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Final Discussion

1. Disclosure Review process
2. Metadata and use of data files
3. Cost of supporting a RDC & funding mechanisms
4. Balancing timeliness with access constraints
5. Communication of research results
6. Access options
7. Project purpose



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Disclosure Review

- Automatic
- Linked data sets pose additional concerns
- Strive to make researchers part of the disclosure review process
- Disclosure review needs to be principle-based
- Distinguish between
  - census/register data
  - Large data sets with high sampling fraction
  - Smaller data collections with high sensitivity
  - Data sets with low sensitivity



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Metadata and use of Data Files

- Consider designating a senior academic as primary authorized user for a given data file as a resource
- Federal agencies need to have a researcher highly knowledgeable about a data set
- Stakeholders need to be involved in options for release strategies for priority data sets
- Data producers need to be involved with research data uses and approval for release of given data sets



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Cost of Supporting a Research Data Center

- Ways to fund a Research Data Center
  - Federal agency supplying the data set
  - Grant from NSF, NIH, NIFA, or other organization
  - User Fees



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Balance Issue of Timeliness with Access Constraints

- Time sensitive research
- Lack of access to federal data may result in use of lower quality data sources
- Approaches needed for data sets formerly issued as Public Use Microdata Files



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Communication of Research Results

- Research results from access to federal data need to be made public
- Special tabulations need to be available to all
- Archives need to be established to facilitate knowledge of related research



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Access Options

- Licensing of researcher with institution
- Research certification
- Balance risk with utility, sensitivity, and cost
- Controls need to be consistent with type of access and research purpose



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Project Purpose

- Issue with users who have nefarious purpose
- Only feasible protection might be researcher certification
- Consider formal ethical approach of a proposed research project
- Equity needs to be a consideration in access protocols



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION

# Follow-up to Sloan Workshop Project

- Presentation today
- Report of project on COPAFS website: [www.copafs.org](http://www.copafs.org)
- Presentation at FCSM 2020 Research and Policy Conference
- Input for individuals to respond to FRN when it is published



*Council of Professional Associations on Federal Statistics*



Alfred P. Sloan  
FOUNDATION