COPAFS

Council of Professional Associations on Federal Statistics

Report on the Discussion at the January 30, 2020 COPAFS hosted Tiered Access Workshop funded by the Alfred P. Sloan Foundation

Welcoming Remarks. Cynthia Clark welcomed participants, thanked the Sloan Foundation for their support, and emphasized that the motivation for this workshop is to help the participants and stakeholders submit comments to the Office of Management and Budget on new guidance issued under the Evidence Act. We were told to expect the Federal Register Notice in mid-summer.

Frameworks – the Five Safes (and others). Julia Lane, NYU
Resource: *Five Safes: designing data access for research.* Desai, Tanvi, Felix Ritchie Richard Welpton, University of Essex

Julia Lane framed her discussion by reminding the group that you cannot assess risk in isolation from quality; both risk and utility are moving targets.

Risk is both increased and decreased by technology.

It is also possible to design access to increase the data utility.

Julia provided the example of the NZ Stats System. Noting that it had started by only producing tables, but now has an integrated data system.

Because of the “skewed nature” of human behavior, the data with the most utility describes small populations, i.e., youth at high risk of not graduating from primary school. It also means that highly sensitive data has a high risk of disclosure. This high- risk nature of the data implies that making safe inputs (safe data from a five safes perspective) that preserve these rare characteristics would be challenging if not impossible. This means that protecting confidential information should focus on protecting **outputs** and not necessarily protecting **inputs.**

Some guidelines suggested by Julia

1. Incentivize compatible approach
2. Tie access to contribution
3. Identify and post utility measures “High quality publication is not the goal – utility is the goal”
4. Administrative records need more meaningful metadata documentation in order to determine their utility
5. Move from ad-hoc determinations to rules-based, and eventually principles based.
6. Engage researchers in part of the process; e.g. help with documentation, etc.

A few reminders

1. This is broader than Title 13 data, and that the decennial census is a very unique use case
2. There are high utility use cases that are beyond the university environment

Organizations releasing data might consider developing a risk mitigation dashboard for data users.

The Safes – Small Discussion Groups

Safe Settings – there is wide agreement that a safe setting does not need to entail a physical, SCIF (Sensitive compartmented information facility like setting.

Key features of a safe setting are
1. IT based solutions, including appropriate encryption,
2. Continuous monitoring, and
3. Physical requirements should be the exemption rather than the rule for a safe setting.

The National Data Standard needs to have the following characteristics

- Vertical
- Scalable
- Applies modern security principles

Some examples of IT based solutions

- Secure log-in
- Internet run through a vpn
- Ability to run security scan
- Biometric identifiers

FISMA levels have been used to help determine which “setting” is appropriate for access with FISMA data. A “high” level designation would be the only situation limited to a physical setting.

Some considerations: Sharing researchers’ intellectual property. i.e., Raj Chetty and Matt Desmond who have put in significant effort to create valuable data sets that are only accessible through their own “labs”.

NCES has successfully used institutional licensing.

Safe Outputs
Three dimensions

- Likelihood of re-identification
- Likelihood of a malevolent intruder using this source
- Degree of harm (sensitivity)

Technology decreasing risk: Increased competitive access to data has reduced risk to federal agencies

Continuum on safe outputs:

- Public data - no output reviewed
- Researcher controlled output (middle tier): researcher manages their own based on data agreement.
- Secure facility/restricted: output reviewed

The following principles were asserted. There was not uniform agreement to the first two principles.

1. Data intruders to statistical output are mythical. (for non-census or non-registry data). Don't optimize to the mythical intruder.
2. The likelihood that complementary disclosure is a threat is vanishing (there are not a sufficient number of malevolent intruders who will review output of statistical agencies.)
3. Pursue automation of a principle-based approach. (Start tracking judgement to use machine learning).

NOTE. Need to collaborate with cybersecurity and criminal experts, including misinformation experts and election officials. In a rules-based disclosure avoidance system, you can't have less than 3 people. In a principle-based system, value judgement says you can override rules and determine the value of output and the system risk.

Safe People

A key consideration in "Safe People" is equity. There is a major danger of restricting access to a small group of elite researchers.

Creating "Safe People"

- The established training protocol should include an ethics consideration with a "people first" perspective
- Recognize that an Affiliation with a recognized research or public institution is a means of ensuring safe people.
- Apply individual and institutional sanctions that include reputational cost (due process required for violations of the agreements).
- Training and sanctions should correspond to the level of data sensitivity

The biggest issue in regard to people relates to unaffiliated researchers who lack institutional standing. Licensing and bonding were suggested but neither provides a place of last resort.

Safe Projects

- Require a sound research design, with ethical considerations (IRB review)
- Take into account the number of data linkages
- Should not be provided for lawful or constitutional uses (i.e. CIPSEA data could not be used in court cases)
- Must be informed by the conduct of similar projects. Useful to have a repository of research with same data set.

Safe Data". Data Confidentiality Classification Tool. Steven Thomas, Statistics Canada
Resource**:** Statistics Canada's Confidentiality Classification Tool

One-size-fits-all access solutions unnecessarily limits researcher access to valuable data resources. Measuring the confidentiality of datasets allows disclosure risks to be mitigated properly and then to be able to take calculated risks with data access.

The CCT is a self-administered tool in which probability of disclosure (attribute, identity, inferential, residual) and severity or harm (severe, high, medium, low, negligible) are combined in a matrix to "score" the level of confidentiality as 1 to 9 where 9 is the highest risk. Alone the CCT score is an awareness tool that allows the data custodian to be aware of the risks associated with their microdata.

The CCT score has been combined with accreditation level (safe People) to determine the "Safe Setting". This can be used for access options for researchers but also for employees where riskier datasets are only accessible within more secure settings...

Governance Board acts as the data custodian, reviews the assessments, and reduces risk when possible. Directors approve. A review committee exists to standardize practices and to hear appeals.

In conjunction with the discussion of the CCT, Steven mentioned that at Statistics Canada there are several different approaches for researchers to gain access to their data files as well as linked data files. The CCT provides a sensitivity rating for the data that is used with these approaches.

- Research Data Centers (physical and remote access to data at the centers)
- Remote system to make requests (Real-time Remote Access).
- Synthetic data sets
- Public Use Microdata Files (PUMFs)

The challenge for Statistics Canada is to evaluate the existing access scenarios and to identify areas where researchers are unable to access data. The goal is to develop new solutions for these cases.

Examining the UK's 3 Tiered Model. Matthew Woollard, University of Essex
Resource: *UK Data Service Data Access Policy*

UKDS assumes open where possible, closed when necessary following these strategies

- Protection of identities
- Processing ground
- Regulated access
- Safeguards and security

Spectrum of Access

- Open – no disclosure risk; **no** license needed.
- Safeguarded – zero to low disclosure risk, authentication, authorization and auditing; certify for research purposes – not prosecuted
- Controlled – disclosure risk/personal data, added safeguards using the Five Safes framework

Five Safes in terms of Access Control

- Safe Settings [Distribute Data => Distribute Access]; commensurate with data sensitivity and researcher qualification,
- Safe Data [Appropriately controlled for disclosure risk]
- Safe Projects [Appropriate use of data]; data management plan required with application
- Safe People [Appropriately trained users]; qualified researcher
- Safe Outputs [Appropriate levels of disclosure control]; review required by data holder

One challenge of PII – it is on a scale - but the law is binary.

Classifying risk – reduction of risk must be acceptable to the data owner. What is "reasonable likelihood of disclosure"?

Tests for reasonable expectations

- Is the intended processing compatible with the consent gained and promises made?
- Implications of wording for future work/prospective studies
- Would the participants have a reasonable understanding of what was going to happen in the future?

Fives Safes enables safe access to data that meet the needs of data protection yet fulfils the demands for open science and transparency

- Safe data - treat data to protect confidentiality
- Safe people - educate researchers to use data safely
- Safe projects - research projects for 'public good'
- Safe settings – Secure Lab environment for personal data
- Safe outputs – Secure Lab projects outputs screener

Tiers or categories are determined by sensitivity of data in combination with project proposal and researcher's qualifications and institutional backing

Different type of access (setting configuration) identified for each tier or category

Closing Discussion. Points made in reviewing the day's presentations and discussions.

Disclosure Review Process.

- The disclosure review process needs to be automated. The number of requests is outnumbering the capacity of organizations to provide timely feedback. It is unclear if there are completely automated vetting solutions. There seem to be automatic review processes but even in those the actual vetting of outputs for confidentiality risks seem to be a manual process.
- Linked data sets complicate the problem of automating the disclosure review process.
- Some comments were made as to how to make the researcher part of the disclosure review process. This would help in two ways. First, the workload for the vetting officer

would be reduced and simplified. Second, there would be less risks if the researcher was fully engaged in understanding, identifying, and treating disclosure control issues.
- Disclosure review needs to be principle-based.
- Need to distinguish between census/registry data, large scale data collections with high sampling fraction (American Community Survey, and smaller scale data collections with high sensitivity, and those with low sensitivity.

Metadata and Use of Data Files

- Useful to designate a senior academic as a primary individual authorized user for given data files where the researcher can use as a resource in learning about the data file.
- Within federal organization there also could be an individual who has insider knowledge to the dataset and is designated as a resource
- Stakeholders need to be involved in options for release strategies for priority data sets
- Data project staff should be involved with research data uses and any approval process for access to restricted data files

Cost of Supporting a Research Data Center (RDC). There is a cost of maintaining data files, providing metadata, performing linkages, doing disclosure review. This needs to be specified.

- Sometimes funded by agency archiving data at the RDC
- Sometimes researcher funded by grant from NSF, NIH, NIFA, etc.
- Data sponsor may put on a user fee

Issue of Timeliness vs Access.

- Some research needs are time sensitive
- Researchers may use state or local data in lieu of less timely federal data
- Agencies need to find an approach to providing some type of public use files for major data sets. Middle ground needs to be determined for major data files

Communication of Research Results

- Some agencies publish special tabs requested by a data user so that others have information
- Research results need to be provided
- Archive could provide summary of research done on specified data sets as well as researcher

Access Options

- Licensing of researcher with institution
- Some type of Researcher certification could be developed for researchers without an institution – access could be at more secure site if required.
- Need to balance risk with utility, sensitivity, and cost
- Controls should be consistent with type of access and research purpose

Project Purpose

- Issue with users who have nefarious purpose.
- Probably can’t protect against them except through researcher certification process.
- Formal ethical approval of research project may be an approach
- Need to consider equity in access protocols

Attendees at the Workshop

Julia Lane, NYU
Amy O'Hara, Georgetown University
Maggie Levenstein, ISR/UMI
Matthew Woollard, UK Data Archives
Steven Thomas, Statistics Canada
Maria Filippelli, Leadership Conference on Civil Rights
Arloc Sherman, DC Center for Budget & Policy Priorities
Claire Zippel, DC Center for Budget & Policy Priorities
Brock Webb, OMB Statistical Policy
Catherine Fitch, University of Minnesota
Tom Krenzke, Westat
Michael Davern, NORC
Steve Pierson, American Statistical Association
Cynthia Clark, COPAFS
Corinna Turbes, COPAFS

Sponsored by

Alfred P. Sloan
FOUNDATION