COPAFS

Council of Professional Associations on Federal Statistics

# Tiered Access Workshop Briefing Book

District Architecture Center
421 7th Street NW, Washington, DC 20004

January 30th, 2020

Sponsored By

Alfred P. Sloan
FOUNDATION

# Background

Late in 2018, Congress passed the Foundations for Evidence-Based Policymaking Act. As a result, the Office of Management and Budget must provide new federal guidance to expand secure access to data acquired under the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). The guidance will establish statistical standards which agencies will use to determine the level of data access that corresponds to the sensitivity of data records. Additionally, the guidance will provide procedures to conduct comprehensive disclosure risk assessments for identified data sets.

Public data users have diverse needs and concerns; it will be necessary for the Office of Management and Budget (OMB) to hear from data users in order to effectively implement access policies and procedures that satisfy public data needs. The Council of Professional Associations on Federal Statistics (COPAFS), with support from the Alfred P. Sloan Foundation, is holding this workshop to make recommendations to OMB that will help inform this new guidance.

The goal of the workshop is to provide practical recommendations that will inform access policies to publicly funded data for research which benefits our society and economy.

## Key Questions

This one-day invited participant workshop in Washington, D. will discuss the topic of tiered access to statistical data. Key questions include:

- What are the general principles that should be used for assessing data sensitivity?
- How many tiers are needed to best meet data user and researcher needs, while addressing privacy concerns?
- What guidance to the federal statistical system will facilitate a transition from current procedures for data release and access to a multi-tiered approach? What primary uses will a multi-tiered approach support?
- What is needed to assist data users and researchers to adopt and transition to a new model?

## Learning Objectives

By the end of this workshop, participants should be able to draft their own informed Federal Register comment to OMB after gaining:

- An understanding of OMB's mandate to provide guidance, including challenges and requirements
- A shared vocabulary and common understanding of relevant concepts and information
- A broader perspective of different stakeholders' needs to help inform the development of OMB guidance
- An introduction to procedures and practices instituted internationally addressing data release and data access.

This Briefing Book includes background on relevant academic work, key terms and definitions, and other materials to help facilitate a discussion during the workshop.

# Contents

Background 2

Agenda (Tentative) 4

Meeting Information 4

Principles for Risk Assessments 5

Critical Concepts and Terminology 5

- Data Assets 5
- Data Privacy vs. Confidentiality 6
- Data Sensitivity 6
- Statistical Disclosure Limitation 6
- Public vs. Restricted Use Data 6
- Microdata 7
- Data Deidentification and Anonymization 7
- Data Quality 7
- Data Lifecycle (at Stats Agency/Unit) 8
- Risk-Based Approach to Management 9
- Risk-Utility Confidentiality Mapping 9
- Data Sensitivity Levels 10
- Tiered Access Levels 10
- Access and Distribution Modes 11
- Differentiating Sensitivity. Access Tiers, and Access/Distribution Modes 11
    - Additional Resources: 11

Existing Frameworks 12

- DoD Security Impact Levels 12
- HIPPA 12
- Five Safes 12
    - Five Safes Framework Examples 13
    - Five Safes Access Solutions 13

Recommendations from previous COPAFS Workshop 15

## Agenda (Tentative)

| | |
|---|---|
| 8:30 – 9:00 | Registration Opens (Coffee and pastries available) |
| 9:00 - 9:15 | Welcoming Remarks - Cynthia Clark, **COPAFS** |
| 9:15 - 10:00 | Frameworks – the Five Safes (and others)<br>**Resource**: Five Safes: designing data access for research, Desai, Tanvi, Felix Ritchie Richard Welpton, University of Essex |
| 10:00 - 11:00 | The Safes – Small Discussion Groups |
| 11:00 – 12:00 | Report on Discussions to Workshop Group |
| 12:00 -1:00 | Lunch (on site) |
| 1:00 - 2:00 | "Safe Data" – StatCan's Data Confidentiality Classification Tool<br>**Resource:** Statistics Canada's Confidentiality Classification Tool |
| 2:00 - 3:00 | Examining the UK's 3 Tired Model<br>**Resource**: UK Data Service Data Access Policy |
| 3:00- 4:00 | Tiers |
| 4:00-4:30 | Closing Discussion |

## Meeting Information

Location: District Architecture Center, 421 7th Street NW, Washington, DC 20004

Metro:

- Judiciary Square (Red line) *0.4 miles*
- Archives/Navy Memorial (Green and Yellow) *0.1 miles*
- Metro Center (Red, Blue, Silver Orange) *0.4 miles*

Parking:

There is plenty of paid parking available near the venue. The closest available parking is at *616 E Street.*

# Principles for Risk Assessments

Replacing the traditional binary classification of statistical information and expanding access will increase the utility of data assets and modernize the Federal Statistical System. The focus is determining the level of data access that corresponds to the sensitivity of data records. The goal is to expand access to data for evidence building.  The plethora of information today and its online availability is driving the need for an effort to provide increased confidentiality protection to individual and establishment data. In addition, we will build a common framework for making uniform access decisions using the Five Safes Framework.

New capabilities in technology such as online data tools with built-in disclosure protections, virtual enclaves in secure cloud settings, and secure sites will facilitate this effort.  These new capabilities are contributing to safer ways to expand access to restricted data sets beyond the traditional binary system of 'public' or 'restricted' data. This effort must include enhancements to traditional Statistical Disclosure Limitation (SDL) methods and more uniform risk-based management principles. Switching to a risk-based approach where specified criteria are used to assess the sensitivity of the data records can facilitate an appropriate selection of data access modes.

There are five key principles that must be accepted in order to release useful data while managing risk.

1. ***Risk is managed or mitigated, but never eliminated.*** There are no absolutes in providing privacy and confidentially protections. The goal is to manage and control known risks to the greatest extent possible and implement best practices at the time the decisions are made.
2. ***It is impossible to have all the facts when making decisions about making data safe.*** There are an unknown number of events, technologies, or other temporal factors that will require adjustment and changes in the future that we will not be able to anticipate beforehand.
3. ***Privacy protection is a limited resource*** that continues to decline over time as more data are released. The process of disclosure review is highly contextual and involves not only the data in question but all data available (past, present, future) that could be used in combination.
4. ***It is an inefficient use of resources to overprotect data*** with excessive controls and/or an overabundance of protection measures that render the output unfit for the use it was created and approved for.
5. ***The goal is to manage data and access at the least restrictive level,*** allowing access and maximizing distribution modes for approved uses. In doing this, we seek an optimal balance of data quality, risk, and access to provide maximum value and use of federal data.

# Critical Concepts and Terminology

This section introduces several key terms and concepts used throughout.

## Data Assets

Data assets are defined as, "A collection of data elements or data sets that may be grouped together."[1] This definition means many different things throughout the data lifecycle. For the purposes of this regulation, a data asset is: (1) one that has been identified by the Chief Data Officer (CDO) or other authority for the statistical agency or unit's data inventory; (2) identified or described by a requestor for evidence building, including assets that are derived from use of automated statistical disclosure limitation (SDL) tools; (3) a new asset that will be included in the inventory at a later date; or (4) unanticipated or un-resourced data assets identified for use by external parties including assets that may require additional support or funding to realize.

[1] 44 U.S.C. § 3502(17)

## Data Privacy vs. Confidentiality

Privacy and confidentially tend to be used interchangeably in many authoritative sources. However, here the two are treated as distinct. The most basic distinction is that privacy is from the viewpoint of the respondent and confidentiality applies to the data collected. The respondent's privacy are legal rights to control or consent to what data is available and how it is collected. Confidentiality refers to how the data collector promises to protect and use the respondent's private data while only sharing what is authorized. Collecting data under a 'Pledge of Confidentiality' refers to the legal promises being made to protect the data and penalties rendered upon the collector for violation of the agreement. That is the context or basis for use of the terms privacy versus confidentiality throughout this document.

## Data Sensitivity

Data sensitivity is a measure of the potential harm resulting from re-identification. This harm to a respondent (person, business, etc.) can be in the form of, but not limited to, the physical, mental, financial, legal, discriminatory, trust, or reputational damage resulting from disclosure of their information.  Evaluating data sensitivity is not new to the statistical system. Substantial harm, embarrassment, inconvenience, or unfairness were important factors when establishing safeguards to maintain confidentiality.[2] There were expanded in the 1977 Privacy Protection Study Commission report[3] which also recognized differences in data sensitivity. The 2007 CIPSEA guidance requires that sensitivity be considered in collecting. maintaining, using, publishing, and providing access to data under a pledge of confidentiality.[4] The regulation intends to provide criteria to assess the sensitivity level of a data set or item.

## Statistical Disclosure Limitation

Statistical Disclosure Limitation (SDL) is the process of balancing data quality (fitness for use) and disclosure risk to make data assets safe for release. Agencies must use current best practices that minimize the risk of disclosing confidential data from statistical products.[5] Many of the SDL methods and procedures can be found in the Federal Committee on Statistical Methodology's *Report on Statistical Disclosure Limitation Methodology*.[6]  This report is currently being substantially revised.

## Public vs. Restricted Use Data

Public-use data are those where the risk of identifying individuals or individual entities has been reduced to an acceptably low level. In order to create a publicly releasable data asset, SDL techniques are performed to generate a Public Use File (PUF).

Restricted-use data are produced to when data or level of detail not found in the PUF is needed for approved research and evidence building purposes. To maintain confidentiality, access to restricted-use data is constrained to authorized agents (e.g., approved researchers), with additional constraints on where and how the data may be accessed and analyzed. Additionally, agency review of analytic results is done to control and prevent the release of confidential or private information. Restricted-use data encompasses the principle of least privilege model where researchers only have access to the data they need for their approved project.

[2] The Privacy Act of 1974, 5 U.S.C. § 552a. https://www.govinfo.gov/content/pkg/USCODE-2012-title5/pdf/USCODE-2012-title5-partI-chap5-subchapII-sec552a.pdf

[3] Privacy Protection Study Commission. (Jul 12, 1977). https://www.ncjrs.gov/pdffiles1/Digitization/49602NCJRS.pdf

[4] i.d. 2007 CIPSEA

[5] *79 FRN 71610* at 71611

[6] (WP 22) Report on Statistical Disclosure Limitation Methodology, (Revised 2005), 1994 (WP 22). https://nces.ed.gov/FCSM/pdf/spwp22.pdf

## Microdata

A microdata file consists of individual records, each containing values of variables for a single person, business establishment or other unit[7] [e.g. respondent]. Microdata may contain very detailed and/or specific information which can be used either directly or indirectly to reidentify the respondent. Direct identifiers are things such as name, address, social security number, etc. that refer to an individual respondent. Indirect identifiers are variables when used in combination with other data sets may reveal enough information to directly identify a respondent. Indirect identifiers might be geographic location, schools attended, dates of important events, memberships in specific organizations, etc.

## Data Deidentification and Anonymization

Data deidentification, also known as anonymization, is the process of removing, masking, or encrypting data elements that can be used to directly or indirectly identify a respondent. There are many ways of releasing data assets such as aggregating or grouping together several respondents. The idea is that the respondents are averaged over a wide enough range that reidentifying a single individual would be extremely unlikely. An example of this might be: what is the average household income in the zip code 12345. Aggregating data is useful but limits the usefulness of the data for in-depth research. In order to maximize the utility of the data, it may be necessary to create a restricted data set accessible to authorized individuals under conditions that mitigate the risk of reidentification. Applying appropriate SDL methodologies allows a statistical agency to promise confidentiality while also facilitating a restricted release of information based on the underlying data.

## Data Quality

Data quality as defined in the Information Quality Act[8] (IQA) means quality, objectivity, utility, and integrity of information.[9] All statistical agencies strive to collect, process, and produce the highest quality data possible consistent with the IQA, OMB Statistical Directive 2, and agency quality guidelines.[10] The IQA requires agencies to conduct pre-dissemination review of their information products.[11] During this review, each agency should consider the appropriate level of quality/detail for each of the products that it disseminates based on the likely access to that information.[12]

Reducing data quality or detail does not mean the data are bad or no longer of good quality. Quality is multidimensional and various aspects can be adjusted to optimize for intended use access. One dimension of data quality is whether it is useful for the analysis being performed. As seen in Figure 1 below, the utility dimension of data quality is plotted against privacy protections. The ideal situation is to have both maximum utility and privacy, but that is not possible. Since confidentiality is a requirement, that risk must be kept to a low, acceptable level. Therefore, choices are made to reduce the data quality (utility) to a point where the trade-off is acceptable.

[7] i.d. WP 22

[8] Information Quality Act. (2002). 67 FR 5365. https://www.govinfo.gov/content/pkg/FR-2002-02-22/pdf/R2-59.pdf

[9] Also see intermediate ICSP data quality product (Obtain more info reference)

[10] SPD#1 https://www.govinfo.gov/content/pkg/FR-2014-12-02/pdf/2014-28326.

[11] Guidelines, sec. III.2, 67 FR at 8459 ("As a matter of effective agency information resources management, agencies shall develop a process for reviewing the quality (including the objectivity, utility, and integrity) of information before it is disseminated. Agencies shall treat information quality as integral to every step of an agency's development of information, including creation, collection, maintenance, and dissemination. This process shall enable the agency to substantiate the quality of its information through documentation.

[12] Improving implementation for the Information Quality Act, OMB (2019) M-19-15

A data asset may not have a PUF if the ‘fitness for use’ cannot be achieved because the amount of quality loss to make it releasable would essentially make it unusable. In that case, the data are only accessible in a protected or restricted mode to ensure the quality and fitness for use is acceptable.

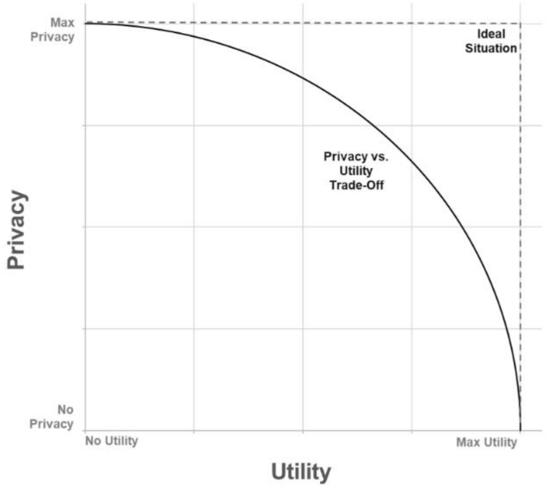

Figure 1. Data utility versus privacy.

As part of the metadata, it is important to document what adjustments, if any, were made to the data quality or detail. Adjusting the level of data quality, for instance, might involve coarsening data by combining response categories and removing sensitive variables, thus resulting in some information loss. By adjusting the data quality, the outputs can result in more accessible data assets at different levels of sensitivity.  Regardless of output, it should be apparent that making data useful means accepting some amount of risk of re-identification to achieve utility or benefit.

## Data Lifecycle (at Stats Agency/Unit)

The process of producing data assets is generalized in Table 1 below. Every statistical agency or unit will have some variation depending on their mission, products, and data customers.

*Table 1. Generalized movement of data (data lifecycle) at a notional statistical agency*

| Movement of Data at a Notional Statistical Agency | | |
|---|---|---|
| 1 | Ingest | • Collect survey data<br>• Acquire administrative data<br>• Obtain/purchase data from a third party (e.g. commercial data) |
| 2 | Process | • Prepare data for delivery<br>• Create new file without PII |
| 3 | Provision | • Deliver data to production/research servers |
| 4 | Use | • Produce safe output<br>• Create comingled files |
| 5 | Destroy | • Destroy original files |

Data can be acquired from different sources and linked together during processing. Processing can be very complex, as data may contain missing or unexpected values, formatting problems, data duplication, and a variety of other issues that need attention. Typically, before data is used by researchers, PII is masked or removed. One method is to create a Protected Identification Key (PIK) that allows linkage of data assets without revealing any direct identifiers such as Social Security Number. Many different data assets and intermediate products can be generated during use. The goal is safe output where data are released only after SDL and approval by the agency's Disclosure Review Board (DRB) composed of a panel of experts. Data sets for approved research projects are retained based on statutory requirements for records management or per a data use/contractual agreement depending on the source of the data. After the agency or unit has met the obligations for data asset retention and/or no longer requires the asset, it is physically or electronically destroyed using approved methods.

## Risk-Based Approach to Management

Protection of CIPSEA data has been developed according to the principle of disclosure risk, which considers both the probability of an unauthorized disclosure and the expected harm from such a disclosure.[13] The cost in managing risk must be proportional to the utility of the information sought and balanced with the amount of risk incurred, and whether that cost is reasonable.[14]

Standardizing and making the process consistent for all CIPSEA data, including data acquired under the Presumption of Accessibility can continue to minimize risk while finding efficiencies to produce more data and expand access. All data are protected and made available in non-identifiable form, and risk is minimized as much as possible given many contextual and often temporal factors. OMB Circular A-130[15] emphasizes the need to manage privacy risk, which is further explored in National Institute of Standards and Technology (NIST) Interagency Report NISTIR 8062[16] to "Develop an engineering approach to privacy."

## Risk-Utility Confidentiality Mapping

The risk-utility confidentiality map was developed by Duncan, et. al.[17] to seek the best SDL method that minimized risk and maximized utility. The general concept is to understand the continuous relationship between utility and risk such that an optimum point can be found. Figure 1 below is a graphical representation of the risk-utility map.

[13] i.d. 2007 CIPSEA

[14] i.d. 2007 CIPSEA

[15] OMB Circular A-130 Revised. (July 27, 2016). *Managing Information as a Strategic Resource* https://www.whitehouse.gov/omb/information-for-agencies/circulars/

[16] NISTIR 8062. (January, 2017). *An Introduction to Privacy Engineering and Risk Management in Federal Systems. https://doi.org/10.6028/NIST.IR.8062*

[17] Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map*, Technical Report 121. National Institute of Statistical Sciences. Research Triangle Park, N.C. Also, a Los Alamos National Laboratory Technical Report, LA-UR-01-6428. https://www.niss.org/sites/default/files/technicalreports/tr121.pdf

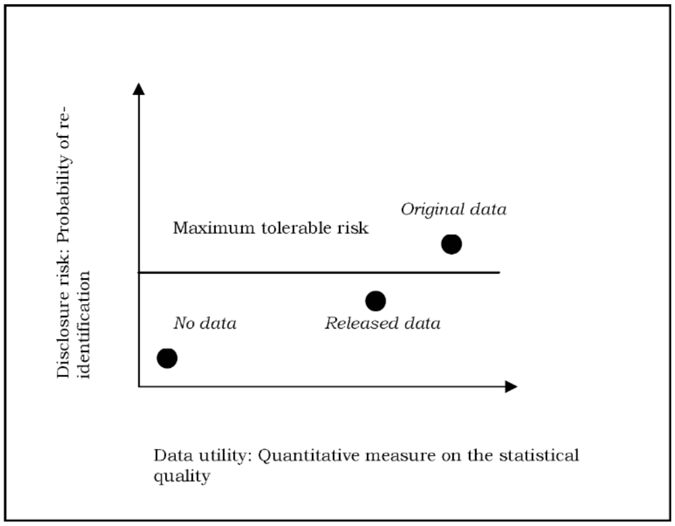

*Figure 1. R-U Confidentiality Map*[18]

It is not always possible to quantitatively compute the risk-utility curve of a particular SDL method. Depending on the application, the lower limit for risk tolerance can be defined based on fitness for use. Utility is only one dimension of data quality and there are other factors that could influence the amount of tolerable risk allowed. With this concept, multiple risk limits can be defined to develop different levels within risk-utility limits.

## Data Sensitivity Levels

Applying controls and safeguards to protect data, the cost of the protections needs to be managed against realistic assessments of risk.  Using available resources appropriately ensures that the maximum amount of data is available and of the highest quality possible within the means of the provider.

Like the risk-utility map, developing data sensitivity levels requires segmenting the quality-risk curve into min-max thresholds that define the acceptable limits for adjustment. Defining the thresholds that delineate different sensitivity levels as well as evaluation of a data asset will be discussed later.

## Tiered Access Levels

Tiered access levels implement the Information Technology (IT) security concept of “least privilege” for allowing access to the systems that collect, store, process, and distribute data. “Least privilege” is giving only access to data and services required to accomplish a job, often with time-bound constraints or limits. For access to restricted data assets, a ‘need to know’ is required to grant access.

The Federal Information Security Management Act (FISMA) required the establishment of a tiered framework for information and IT systems based on the risk and magnitude of harm associated with unauthorized access. The established standard for assessing the FISMA level is accomplished with the National Institute of Standards and Technology (NIST) Risk Management Framework (RMF). The RMF is applied to information and IT systems to determine the level of security required based on the confidentially, integrity, and availability of the data and IT infrastructure. The resulting determination is a FISMA risk rating: high, moderate, or low.

Tiered level access controls ensure that only authorized access by a person and/or machine is allowed. The machine could be a credentialed device, software, server, or service. Ultimately these concepts lead to more granular access control, which is beyond the scope of this discussion. Suffice to say, if a device is

[18] Hundepool, A., et. al., (2010). *Handbook on Statistical Disclosure Control version 1.2.*, https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

considered compromised by the security system, even an authorized user will not be able to use it to gain access to the tier.

## Access and Distribution Modes

The data can live in a single access tier and have multiple access distribution modes (ADMs). Figure 2 below is a general operational picture of how the federal statistical system (FSS) provides data asset access and/or distribution through different modes in 2019.

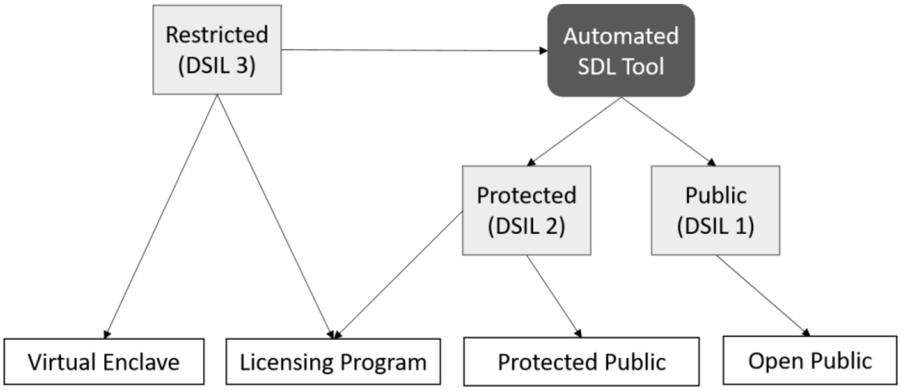

*Figure 2. Single data source, multiple paths to access the data or cleaned output.*

Data products produced (not shown) from using any of the non-public modes must go through SDL and a DRB to be released as open public data. Automated tools can be authorized to act as an intermediary, performing the proper SDL methods automatically to produce different sensitivity level outputs. Restricted data are available at a Federal Statistical Research Data Center (FSRDC), in a virtual enclave or within a user-controlled enclave that conforms to terms and conditions of a use agreement or license. Protected data are offered to run on a user's enclave with or without specific conditions. Protected data are also available through a protected website where they can be downloaded under a license for use. Finally, data can also be distributed as open public data with no use restrictions.

## Differentiating Sensitivity. Access Tiers, and Access/Distribution Modes

Several data and IT concepts are being combined because the data exists within IT systems, and those systems must be appropriate for safeguarding the data. The data sensitivity level, tiered access level, and the access and distribution mode (ADM) are related, but each is distinct and not interchangeable.

Sensitivity level is based on the risk balanced with quality for appropriate protections in accordance with the RMF. The access tier is a combination of IT systems, services, and security controls appropriate for the sensitivity level of the data it hosts and protects. Data are stored, processed, accessed, and distributed in many different ways, or modes.

## Additional Resources:

- Foundations for Evidence Based Policymaking Act
- Commission on Evidence Based Policymaking Website
- Report on Statistical Disclosure Limitation Methodology, (Revised 2005), 1994 (WP 22).
- OMB Circular A-130 Revised. (July 27, 2016). *Managing Information as a Strategic Resource*
- *An Introduction to Privacy Engineering and Risk Management in Federal Systems.* NISTIR 8062. (January 2017).
- Duncan, G., Keller-McNulty, S. and Stokes, S. (2001). *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map*, Technical Report 121. National Institute of Statistical Sciences. Research Triangle Park, N.C. Also, a Los Alamos National Laboratory Technical Report, LA-UR-01-6428.
- Hundepool, A., et. al., (2010). *Handbook on Statistical Disclosure Control version 1.2.*,
- Statistics Canada's Confidentiality Classification Tool

- Public Access to NSF-Funded Research Data for the Social, Behavioral, and Economic Sciences Workshop Report
- Security Module, Coleridge Initiative

# Existing Frameworks

This section contains other frameworks useful as a reference and for background knowledge.

## DoD Security Impact Levels

The United States Department of Defense (DoD) has classified national security information at three different levels (confidential, secret, top secret) depending on the severity of damage to national security if information becomes available. In the early 2010s, the DoD created Security Impact Levels (IL) for public, Controlled Unclassified Information (CUI), and secret data to expand use of commercial cloud computing capabilities, vendor managed data centers, and other internal capabilities or services. These ILs were later aligned to the DoD Risk Management Framework (RMF) to operationalize the evaluation of data for future use.

The DoD RMF aligns with the National Institute of Standards and Technology (NIST) RMF used by the rest of the Federal Government. It was created after the 2013 National Defense Authorization Act required the DoD and the Intelligence Community (IC) to adopt a risk-based approach to modernize security approaches. The NIST RMF was adapted to the DoD RMF to account for the particularities of the IC and the DoD with respect to National Security Systems. The data risk was based on confidentially, integrity, and availability. The impact was based on the harm to national security if the data was exposed or compromised by an adversary. IT systems were authorized and controlled based on the IL of the data. This naturally created different access levels with a variety of types of access and modes of information distribution.

Additional Reading:
Executive Order 12356--National security information
Department of Defense Cloud Computing Security Requirements Guide (March 6, 2017)
Risk Management Framework for DoD Information Technology (March 12, 2014)

## HIPPA

The guidance for Protected Health Information (PHI) under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule Section § 164.514 has two distinct paths for deidentification of a data set. The first is 'Expert Determination' in which experts review and make the determination as to which identifiers are retained or removed. The second is 'Safe Harbor' where a standard set of 18 identifiers ~~that~~ are removed before the data ~~are~~ is released. While both comply with the law, some residual risk of reidentification ~~always~~ remains.

Additional Reading:
Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule Section § 164.514

## Five Safes

The Five Safes framework was developed in the early 2000s and has been broadly adopted by several national statistical offices -- the U. K's Office of National Statistics, Stats New Zealand, Australian Bureau of Statistics, and by Eurostat. It also was developed by New York University's Administrative Data Research Facility (ADRF) which in support of the US Commission on Evidence-Based Policymaking. The Five Safe framework provides a useful tool to assess the disclosure risk of a data set used for a research endeavor.

The Five Safes framework groups all aspects of a research endeavor that uses confidentially protected data along five dimensions:

(1) Safe projects – projects are valuable and appropriate (to the agency's mission);

(2) Safe people – requirements in training, trust levels for access;

(3) Safe settings – data access controls and infrastructure constraints;

(4) Safe data – disclosure risk assessments to determine data sensitivity levels and quality loss associated with different confidentiality protection procedures;

(5) Safe outputs – application of disclosure limitation methods and implementation of disclosure review requirements to assure the confidentiality protection of outputs

## Five Safes Framework Examples

The following are examples of using the Five Safes. These examples are not intended to be an exhaustive list of all the possible permutations but to show how use of the Five Safes could be implemented. The specific attributes of a data set would generate more specific data than some of the general descriptions given.

- Five Safes: designing data access for research, Desai, Tanvi, Felix Ritchie Richard Welpton, University of Essex
- Regulating Access to Data, UK Data Services
- Integrating Data Infrastructure, Stats New Zealand
- Managing the Risk of Disclosure: The Five Safes Framework, Australian Bureau of Statistics
- Eurostat Statistics Explained: European business statistics manual, Eurostat

## Five Safes Access Solutions

### *Restricted Data under License at user Enclave*

A university has a licensing agreement with the data owner to use the data in a protected enclave that is managed at the user's institution. Under the agreement with the university, only sworn CIPSEA Agents may use the data. The data owner may review the data use and the facility to ensure proper use of the data. If the data owner finds violations that may revoke the privileged data use.

- **Projects:** the university manages appropriate projects with respect to the intended data use per the terms of the license agreement
- **People:** all users receive data stewardship training for proper handling and safeguarding of the data. Users sign an acknowledgement of penalties for non-compliance.
- **Settings:** an air-gapped set of computers in a locked laboratory space are maintained by trusted IT staff in accordance with the agreement.
- **Data:** the data owner pre-processes data to mask direct identifiers and to ensure that only the data needed to perform the research is provided.
- **Output:** The data owner's disclosure review board reviews output to ensure it is safe before it can be included in any publicly accessible format.

### *Automated SDL Tool Output to Public Website*

An agency uses an SDL tool that can access restricted data and automatically produces safe output as open data.

- **Projects:** open data, no restrictions
- **People:** all users, no restrictions

- **Settings:** For users, uncontrolled and no restrictions. The agency must maintain the SDL tool, and any other IT assets used in the data supply chain must be properly secured, monitored, and maintained.
- **Data:** Restricted data are only accessible by the SDL tool. For security purposes, the data asset used by the SDL tool should contain no direct identifiers because they are not needed to produce useful output.
- **Output:** The data owners have configured the SDL tool to meet specific criteria for the automated production of open data. Output is validated before it is placed in service.

*Restricted Data at Federal Statistical Research Data Center (FSRDC)*

Agency chooses to use FSRDC to host data and provide researchers tools for analysis. FSRDCs are managed by the U.S. Census Bureau.

- **Projects:** every project is vetted through a process that ensures the data sets requested are appropriate for the research being done, provide a value or benefit to the data producer, and meet other legal use restrictions.
- **People:** researchers must meet all training requirements for data stewardship, complete background investigation, and obtain Census Special Sworn Status.
- **Settings:** FSRDC provides all the physical controls to ensure only authorized researchers gain access to the facility. Computing terminals have been secured in accordance with Census security requirements and allow only virtual access to the data and secure computing resources. Users cannot download or otherwise remove data. Use of a mobile phone or other digital capture device use is prohibited in the facility.
- **Data:** Only the data required are provided. Direct identifiers have been masked with a Protected Identification Key (PIK) to allow approved linkages to other data assets as described by the approved project proposal.
- **Output:** All data owners involved agree to the terms of the disclosure review. Researchers only receive data output that has been reviewed.

# Recommendations from previous COPAFS Workshop

**Recommendation 1:**

We recommend that OMB encourage experimentation among different dimensions and options for access. We recommend that the FSRDC experiment on how it might adapt conditions for access, particularly considering the challenges of users who may not be associated with academic institutions. Promising examples of how this might be accomplished are being done, such as the pilot program on FSRDC remote access. Some key questions to consider in this research are: Why do you have to have a CB employee on-site? Why not cameras observed remotely? Why do researchers have to be Census Specially Sworn Employees? Wouldn't some other contractual arrangement work?

**Recommendation 2:**

The Federal Statistical System requires statistical purpose for access to FSRDC. 'Statistical purpose is defined by CIPSEA as the description, estimation or analysis of the characteristics of groups without identifying the individuals or organizations that comprise those groups. CIPSEA protected data is restricted from regulatory or enforcement use. Uses envisioned by the Foundations of Evidence-based Policy Act for program and policy evaluation, if defined appropriately, could meet the definition of statistical purpose. Any legislative barriers to such uses need to be identified and a plan developed to address them.

**Recommendation 3:**

In designating different tiers of access, a continuum for mitigating risk for each of the five safe constraints -- people, projects, setting, data, and output -- should be developed and applied.

**Recommendation 4:**

Research and development should continue into more promising solutions for access in lieu of public use microdata files, such as synthetic data, and the process for access. One suggestion is to use a researcher passport to streamline the process for applying for access to public-use data files of a type that were previously public.

**Recommendation 5:**

Research better risk quantification/estimation approaches. Westat has done risk estimation dozens of times successfully with at least 3 agencies using statistical modeling in the past 6 years and it produces some helpful information that can be used with other information to make recommendations for placing different products into different tiers.

**Recommendation 6:**

Conduct a holistic review of agency products to attempt to balance data utility with confidentiality and privacy constraints. This would include 1) identifying all the data sources (e.g., surveys), their data products, the main sources of risk, rules of disclosure (e.g., Rule of 3, etc.), confidentiality edits (e.g., cell suppression rules), 2) for tables, determining if the table suppression system can be broken by the product itself or other data products, for a flexible table generator, determine if it is susceptible to differencing attacks, and for microdata estimate the risk. Recommendations can then be made for data architecture and risk mitigation.