



The Researcher Passport: Improving Data Access and Confidentiality Protection

ICPSR's Strategy for a Community-normed
System of Digital Identities of Access

May 1, 2018

A REPORT FROM THE INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

2018 – 01

Authors

Margaret C. Levenstein, ICPSR, University of Michigan
Allison R.B. Tyler, University of Michigan School of Information
Johanna Davidson Bleckman, ICPSR, University of Michigan

Project Team

Margaret C. Levenstein
Allison R.B. Tyler
Johanna Davidson Bleckman
Nyanza Cook

Acknowledgements

This report was prepared with support from the Alfred P. Sloan Foundation and the Institute for Museum and Library Services (RE-01-15-0086-15). Support was also provided by the University of Michigan School of Information's Research Experience for Master's Students (REMS) program, funded by the National Science Foundation. We appreciate the participation of 23 repositories in the study.

Recommended Citation

Levenstein, M.C., Tyler, A.R.B., Davidson Bleckman, J. 2018. The Researcher Passport: Improving Data Access and Confidentiality Protection: ICPSR's Strategy for a Community-normed System of Digital Identities of Access. *ICPSR White Paper Series No. 1*. Ann Arbor, MI: University of Michigan Inter-university Consortium for Political and Social Research.

Executive Summary

Research and evidence-building benefit from the increased availability of administrative datasets, linkage across datasets, detailed geospatial data, and other confidential data. Systems and policies for provisioning access to confidential data, however, have not kept pace and indeed restrict and unnecessarily encumber leading-edge science. One series of roadblocks can be smoothed or removed by establishing a common understanding of what constitutes different levels of data sensitivity and risk as well as minimum researcher criteria for data access within these levels.

This report presents the results of a recently completed study of 23 data repositories. It describes the extant landscape of policies, procedures, practices, and norms for restricted data access and identifies the significant challenges faced by researchers interested in accessing and analyzing restricted use datasets. It identifies commonalities among these repositories to articulate shared community standards that can be the basis of a community-normed **researcher passport**: a credential that identifies a trusted researcher to multiple repositories and other data custodians.

Three main developments are recommended. First, **language harmonization**: establishing a common set of terms and definitions – that will evolve over time through collaboration within the research community – will allow different repositories to understand and integrate shared standards and technologies into their own processes. Second: develop a **researcher passport, a durable and transferable digital identifier** issued by a central, community-recognized data steward. This passport will capture researcher attributes that emerged as common elements of user access requirements across repositories, including training, and verification of those attributes (e.g., academic degrees, institutional affiliation, citizenship status, and country of residence). Third: **data custodians issue visas** that grant a passport holder access to particular datasets for a particular project for a specific period of time. Like stamps on a passport, these visas provide a history of a researcher's access to restricted data. This history is integrated into the researcher's credential, establishing the researcher's reputation as a trusted data steward.

Table of Contents

Executive Summary	2
Table of Contents	3
Introduction	4
Problem Statement	4
Needs Statement	6
Research Methodology	8
Current Status of Restricted Data Access Procedures	9
Data Access Process	9
Conflicting Terminology	10
<i>Repository staff on how they define restricted data</i>	11
Access Methods	12
Training	13
Violations and Consequences	14
Restricted Data Use History	15
<i>Perspectives on importance of prior data use history</i>	16
Solutions	17
Language Harmonization	17
Shared Understanding of Data Security	18
A Durable and Transferable Digital Identifier	18
<i>Professional Roles, Demonstrated Trust, and Track Record of Responsible Use</i>	18
<i>Training Requirements</i>	19
Overview of the Researcher Passport and Visa Processes	19
The Researcher Passport and Visa Provision System: A Roadmap	20
<i>The Researcher Passport</i>	21
<i>Establishing Security Level of the Data</i>	22
<i>Visa Application and Issuance</i>	23
Building the System	25
Benefits	27
Next Steps for the Credentialing Project	28
New Restricted Data Use Agreement Template	28
Training Evaluation and Certification	28
Outreach and Testing	28
Timeline	28
Conclusion	29
Appendices	30

Introduction

Problem Statement

The increased availability of data on individuals, communities, businesses, and other entities, along with increased capacity for data storage and processing, significantly raises the risk of re-identification of individuals or entities who currently presume anonymity. As a result, the data from many traditional surveys that would previously have been made public are now accessible only with significant hurdles for the researcher and the data provider. The risk of disclosure affects the willingness of potential data providers to share data, the ability of researchers to undertake creative and valuable scientific analyses, and the training of the next generation of empirical scientists.

In addition to traditional survey data collected with the explicit consent of research subjects, there is growing recognition of the research potential of administrative and non-designed, organic data collected for non-research purposes. Many repositories lack the capacity to provide the required level of security and protection for these data, or their policies and procedures lack the agility necessary to adapt to the changing landscape in order to accommodate such data responsibly. New models for data dissemination, such as the establishment of multiple administrative data facilities, create new opportunities for research, especially if those data can be linked or analyzed in tandem, but require shared understanding of data security and management and consistent access procedures across facilities. This project recommends standards that, if adopted by multiple repositories, will facilitate commingling and parallel analyses of data from different data custodians.

The growth and recognition of administrative and non-designed, organic datasets, as well as the diversity of data sources, provide new opportunities for creative analyses of multiple datasets across repositories and disciplines. The systems and policies for provisioning access, however, have not kept pace with this growth and indeed restrict and encumber leading-edge science. A number of roadblocks – including differing or inconsistent terminology, varying Internal Review Board (IRB) and legal obligations, and varying and cumbersome application procedures and requirements – can be eased or removed by establishing a shared understanding of data risk levels and minimum researcher criteria for access within these levels. With broad adoption, new and harmonized community norms can eliminate duplicate and inconsistent efforts, wasted resources, and unnecessary risks both to data security and to public trust in the research process.

This report identifies the significant challenges faced by researchers interested in accessing and analyzing restricted-access datasets and proposes the creation of a researcher passport system, allowing credentials to be described and transferred between repositories and other data custodians. It describes recently completed examination of practices at 23 data repositories in the United States, Europe, and Australia, to detail the array of

policies, procedures, practices, and norms that shape the current landscape of restricted data access and outline the necessary steps to articulate and align community standards.

These and related issues have been considered and addressed to varying extents before. The Commission on Evidence-based Policymaking, in its 2017 final report,¹ advocated the creation of a harmonized system for researcher access to administrative data from across federal agencies.² The Dataverse Project at Harvard University is building DataTags,³ a tool to allow for partial automation of dataset risk assessment. Tags can be attached to a file that capture recommended access requirements, required user credentials, and legal terms under which the data must be used.⁴

The Data without Boundaries (DwB)⁵ project in Europe and the U.S. Census Bureau's Federal Statistical Research Data Centers (FSRDC)⁶ model have grappled with this issue. The DwB project resulted in a web-based data request and access system, allowing researchers to simultaneously apply for access to manipulate, extract, and analyze online data from a number of European countries. Challenges such as metadata collection, record linkage, confidentiality protection methods, resource discovery, and software development were researched, documented, and, in some cases, addressed. The FSRDC model, with 29 active data centers around the United States, has successfully addressed the inherent tension between the need to protect the federal statistical system's confidential data and the recognition of the utility of these valuable microdata in advancing research for the public good. The FSRDC model meets the legal requirements for microdata collected by the U.S. Census Bureau, but is more restrictive than necessary for most data custodians that seek to disseminate data responsibly across a wider array of risk levels and most especially data with lower risk. A community-normed system that can flexibly consider the assessed risk level of the data; the experience, training, and stewardship history of the researcher; and the risks inherent in the mode(s) of data access can better moderate risk and can be trusted and applied beyond a single repository.

This paper recommends and describes three key steps. The first is language harmonization. The second and third recommendations are to develop a durable and transferable digital identifier, to act as a researcher passport

¹ The Promise of Evidence-based Policymaking: <https://cep.gov/content/dam/cep/report/cep-final-report.pdf>

² The United States Census Bureau Data Repository at ICPSR preserves and disseminates survey instruments, specifications, data dictionaries, codebooks, and other materials provided by the U.S. Census Bureau. See <https://census.icpsr.umich.edu/census/static/about>

³ DataTags: <https://datatags.org/>

⁴ This project is also implementing differential privacy tools to determine how much of a study's "privacy budget" is consumed by the release of any particular statistic or dataset, to help evaluate the privacy cost of increased data access in any particular instance. The privacy budget is considered the maximum cumulative acceptable disclosure risk level across all analyses conducted on a dataset for a given research project.

⁵ Data Without Boundaries: <http://www.dwbproject.org>

⁶ Federal Statistical Research Data Centers: <https://www.census.gov/fsrdc>

issued by a central, community-recognized data steward, accompanied by a system of visas issued by individual repositories to permit access to particular datasets for a specified period of time. Outlined below are recommended practical steps for identifying and adopting existing training programs, mapping, in matrix form, the interaction between data access method, data user credential level, and assessed data risk, as well as the design and support of underlying technology. These shared terminologies and structural developments will foster a culture of confidentiality and shared responsibility that will better serve the broad social science research endeavor.

Needs Statement

In the current environment, restricted data are often available to the research community only after a lengthy and complicated application process. Our analysis of the application process for 23 repositories finds that these processes are inconsistent, not only in what they require of researchers but even in how they define restricted data, modalities of access, and responsible and trusted data users.

This process usually requires the interested researcher to address the following:

- Detailed data request: The applicant must indicate the desired restricted datasets and variables and any additional data that will be used in the analysis.
- Research topic and plan: The applicant is usually required to provide an analysis plan explaining why the restricted data are necessary to complete the study. Typically the plan must be crafted prior to seeing the data, and sometimes prior to seeing even minimal metadata. While encouraging researchers to commit to a research plan may prevent statistically unsound analyses, it is particularly challenging for analyses of administrative and other organic data where the link from the data to the social science concept or the completeness of the data themselves is not always clear.
- Computing environment and data security plan: Restricted data requests often require that the applicant describe a particular computing environment that the researcher or the researcher's institution provides. In some cases prospective data users are limited to a specific computing environment that the data provider hosts (e.g., a virtual data enclave). A required data security plan specifies the rules, process, and location for accessing and analyzing data (e.g., required IRB review, researcher training, and researcher and institutional safeguards).
- Principal investigator and research team: In many cases, data providers request information about university or other status (e.g., distinguishing between faculty and graduate students; citizens or residents of a particular country; researchers, journalists, and commercial entities; researchers at institutions with Institutional Review Boards; those subject to subpoenas or Freedom of Information Act requests; and those with the legal ability to submit to the requirements of the data custodian).

The conditions under which researchers access data depend on the interaction between data and researcher characteristics. This process is burdensome both for those who try to make data available and for researchers trying to use data. It creates opportunities for people to hoard data and refuse to share, under the guise of protecting confidentiality, or to claim quite legitimately that it is simply too costly to share data safely. This situation becomes worse when researchers try to link datasets and use administrative and other organic or non-designed data.

To that end, this environmental scan analyzes the written policies and other documentation from 23 repositories and interviews with representatives from 10 repositories. It identifies key dimensions of consensus as well as points of difference among repositories that, in order to provide a more streamlined-yet-secure access process, require standardization. We propose a shared system for a researcher “passport” for which individual repositories or data custodians will issue “visas” that provide access to particular datasets under particular conditions for a specified length of time.

Research Methodology

Twenty-three data repositories and service providers around the world were evaluated during this research study. While the majority of repositories are based in the United States, two each are from the United Kingdom, Germany, and Australia, and one from the Netherlands. The international repositories were included to provide a global perspective on the question of research data sharing and to include work done in those countries on similar issues of restricted data sharing and credentialing at organizations like the UK Data Archive and New York University’s Center for Urban Science and Progress (CUSP). Documentation from repositories and screenshots of relevant information from their websites were collected and the text was analyzed using the NVivo qualitative data analysis tool. The document collection encompassed the following:

- “About” pages
- Access control policies
- Access request forms/guides
- Data sharing policies
- Data use agreements
- User guides/Frequently Asked Questions
- Codes of Conduct
- Training materials
- Computing environment requirements
- Legal/ethical policies
- Terms of use
- Government regulations

In addition, interviews were conducted with representatives of 10 of the repositories. The purpose of these interviews is to provide clarity on repository practices for credential development, data security, and access controls.

A total of 355 pieces of documentation were coded using a grounded theory approach to qualitative code-set development. The codes used in the analysis related to the repository’s data access process, computing environment, security, contracts, credentials, technology requirements, data documentation, purpose, secondary users, and required training. The results of this initial document coding informed the criteria for the restricted data access credentialing system.

Current Status of Restricted Data Access Procedures

Over the course of the environmental scan, several things became clear. First, the processes used by the different repositories, while following the basic pattern of “receive application, review, approve or deny access to data,” do not make use of a standardized decision-making process. Second, the repositories do not classify and define data by security and privacy requirements in a standardized manner. Third, while there exists general consensus that data researchers should be trained in handling restricted data, there are no standardized requirements for completing, tracking, and sharing the records of that training or even what constitutes the essential elements of such training. Fourth, there are inconsistencies in how repositories approach data use violations, including both disclosive events and non-disclosive data handling practices, as well as in the consequences for those actions. Finally, there is currently no process to share histories of data use – what datasets were made available, at what level of security, the researcher’s data stewardship contributions, and any history of disclosive or other data handling violations – between data repositories.

Data Access Process

The processes for requesting restricted data differ at every examined repository. The differences fall into the following categories:

- Identity of reviewers and decision-makers
 - Original data producer or provider
 - Repository
 - External scientific and ethical reviews
- Criteria for review
 - Scientific merit
 - Necessity of confidential data
 - Consistency with repository purpose (both legal and community purposes)
- Approval process
 - Same or different processes for data of different security levels
 - Complexity of approval chain (e.g., simple reviews have 1-2 reviewers; complex reviews have more)
 - Sequential or concurrent reviews
 - Length of review process (e.g., decision within days or weeks/months/years)
- Researcher identity verification
 - Cursory (validate affiliation through website searches)

- In-depth (background checks and security clearance requirements)
- Institutional
- Repository account creation
 - Required for data access request submission or not required
 - User-created or repository-created accounts
- Assumption of liability
 - Individual
 - Institutional

There is variation in the standardization of the application process itself within repositories. There are repositories in which the process differs with every dataset requested, based on memoranda of agreement between the repository and the data providers.

Conflicting Terminology

Restricted data repositories do not use the same terminology to identify data along a low-moderate-high security spectrum. For the purposes of classifying practices at study repositories, we make the following distinctions: low security data include no identifiable information or other security-related reasons for restricting access; moderate security data require application, may require a data use agreement, and require repository approval; and, high security data require depositor/data producer approval, a government-issued clearance, and/or secure computing environments. At each of these levels, repositories use multiple terms to identify the security necessary for access to the data, and, between the levels, the same terms were used to indicate different levels of security. For example, at three repositories, “restricted” referred to data which required only moderate security, while at six repositories “restricted” indicated much higher security requirements. See Table 1 below for descriptions of terms used across repositories.

TABLE 1. TIERS, THEIR USE, AND THEIR TERMS

	Most common term at this tier	# of repositories using this term	# of other terms to define this tier	Other terms	Clarifications
Security Tier	Low				
	Public-Use	5	6	Open access, green, data available, open access for registered users, unrestricted, category (cat) 0	Includes data that require repository account or login
	Moderate				
	Restricted	3	5	Yellow, cat A, limited data set, secure dissemination, scientific use, general	Includes data with requirements for applications, DUAs, repository approval, etc.
	Controlled Access	3			
High					
Restricted	6	12	Confidential, government regulated, red, cat B, cat C, restricted release, research identifiable, virtual enclave, physical enclave, closed, special, other access, safeguarded	Includes data requiring depositor or producer approval, government-issue clearance, secure computing environments, etc.	

Most of the data held in these repositories came from social science surveys, but also included are longitudinal health data, external sensor data about urban traffic patterns, genotype and biomarker data, proprietary commercial data, and administrative labor market data, among others. Despite this variation in data types, the motivations behind how the data are classified are similar, even though the classifications themselves are different. During the interviews, repository representatives were asked to describe how their institutions defined “restricted data” for their data holdings. The reasons for data restriction fell into three categories: data provider requirements, national legislation or legal or commercial protections, and a sense of moral duty to protect research subject privacy (expressed without reference to legal or contractual requirements).

Repository staff on how they define restricted data

By data provider requirements:

- *“Fundamentally, the way we would classify most of our content is restricted. Perhaps more accurately described as mediated access.” (SF_001)*
- *“It’s driven by whose data is it, what are the strictures around it, what is the legal framework around it, what’s possible.” (SF_003)*
- *“I would say most of the data that we have is not restricted, or it’s restricted, but it’s more restricted because somebody owns it and doesn’t want to give access to the data to everybody.” (SF_005)*

By national legislation:

- Restriction based on data provider requirements
- *“The data that we hold at the institute, in German law they are called ‘Sozialdaten,’ so social data. ... And there are very, very strict rules for what we have to do with social data and who can get access to social data and for what.” (SF_004)*

By a sense of moral duty:

- *“So, for my approach most of ours is going to be disclosive data, so respondent disclosive data. ... So, restriction for me has to do with your ability to identify a single respondent... And I make a distinction in my mind between like restricted data that are disclosive and data that are sensitive.” (SF_008_01)*
- *“Ours are also, I think, heavily to keep people from being identified if we think that you can still get a unique profile of somebody within the data, also knowing how to recompile the original sample, if you can guess that person could be in the original sample. Because data that are sensitive can be, on sensitive topics, can be made available if you think that people can't be re-identified.” (SF_008_02)*

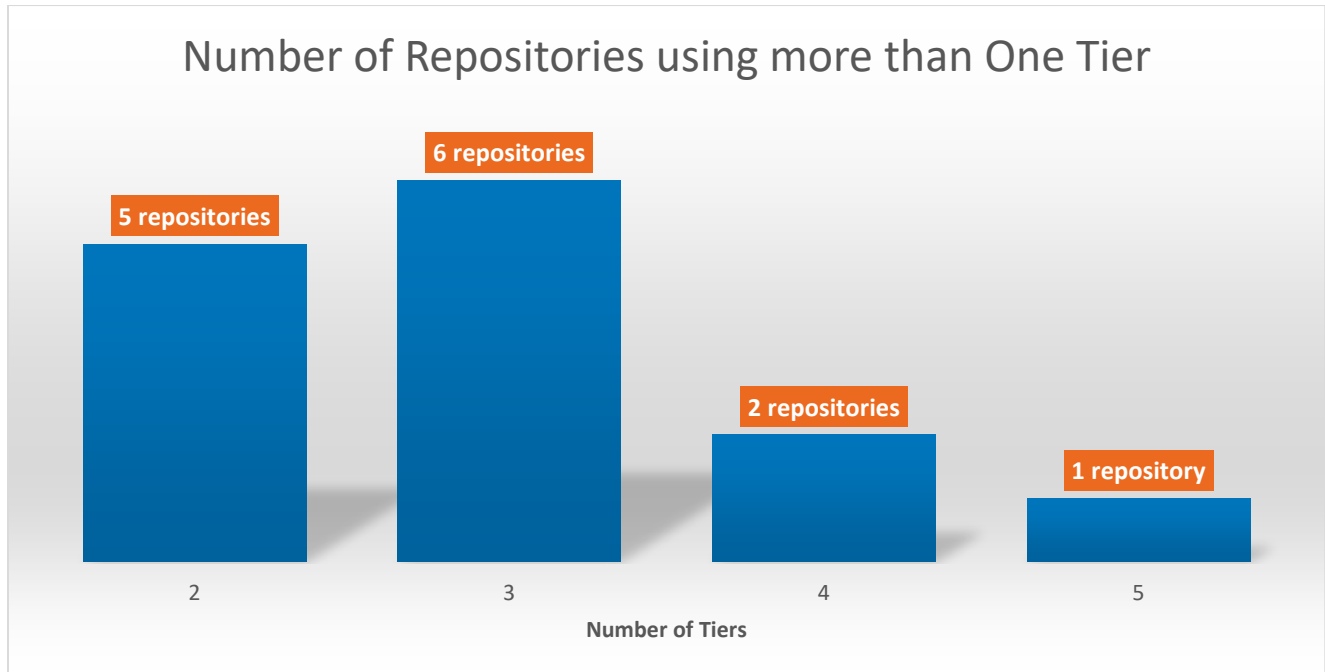
Access Methods

The repositories used similar methods of access for restricted data. Several repositories had multiple access methods based on the sensitivity of the data (see Figure 1 below). These methods are:

- Secure download (through FTP, Dropbox, or other encrypted dissemination) with no restriction on the computing environment
- Secure download to secure local computing environments (e.g., standalone computers in locked offices or secure computing spaces controlled by the user’s institution)
- Online analysis tools (e.g., programs that produce statistical output but do not permit the researcher to view the underlying data)
- Virtual data enclaves (i.e., remote access to a secure computing environment controlled by the repository)
- Secure physical enclaves in which the physical and computing environments are controlled by the repository

For repositories with multiple access methods, required information about the potential data user and intensity of the project review process increased in accordance with the access method security level.

FIGURE 1. TIERS & REPOSITORIES



Training

U.S. federal research funding agencies – National Institutes of Health (NIH), National Science Foundation (NSF), and National Institute of Food and Agriculture (NIFA) – require those who receive research funding to complete responsible conduct of research training; only the NIH has stated explicitly what that training should include. As many of the repositories in this study are affiliated with universities receiving such funds, repositories may assume that academic researchers have completed training that meets federal requirements. However, the reporting and explicit requirements for researcher training in responsible conduct of research, proper data security practices, and data management vary substantially across repositories. The breakdown of training requirements among the repositories is as follows:

TABLE 2. REPOSITORY TRAINING REQUIREMENTS FOR RESTRICTED DATA REPOSITORIES

	Require Training to Access Data	Require Training for Account Creation	Require Training in accordance with Data Owner Requirements	Do Not Require Training	Unknown Requirements
# of Data Repositories	6	2	2	6	9

There is no general consensus on either the substantive content or a recommended provider (such as PEERS⁷ or CITI⁸) of data security training across repositories. Many institutions use internal or proprietary training videos and programs that are not shared with other institutions. The training curricula that are required and the number of repositories that require them are:

- Data privacy and confidentiality (4)
- Responsible data use (2)
- Information security (2)
- Enclave access (2)
- Disclosure control (2)
- Data stewardship (1)
- Title 26/Title 13 (1)
- CIPSEA (1)
- Behavior for IT systems (1)
- Import/export of files and analysis applications (1)

Based on these discrepancies, ICPSR will conduct a further review of training materials and requirements in 2018 to provide more detailed and explicit recommendations for training.

Violations and Consequences

The process by which repositories respond to data use violations also requires reconciliation. Repositories often have requirements that include a timeframe for notification of a disclosure event or a violation of the terms of data access (by the researcher to the repository). They may also require notification to an original data provider. These reporting requirements are often, but not always, included in data use agreements signed by researchers and their home institutional signing officials, as well as in any other repository user guides and handbooks. These explicit requirements focus on disclosure events (regardless of intention) but exclude other types of poor data management, such as failing to lock a computer when stepping away from the desk or not resetting an alarm within a secure facility. The enforcement of consequences for either type of offense has not been consistently applied. We need, first, to identify these other inappropriate behaviors, and, second, to create more standardized processes for enforcing consequences at the individual- and institutional- levels. Only with

⁷ Program for Education and Evaluation in Responsible Research and Scholarship: <http://research-compliance.umich.edu/glossary/peers>

⁸ CITI Program: <https://about.citiprogram.org>

consistent application of penalties, for both “major” and “minor” offenses, can consequences be valuable components of a researcher credential.

Consequences, when implemented, can be applied at two different levels, which provide multiple incentives to handle restricted data properly. Potential consequences can include:

- Individual consequences
 - Legal consequences
 - Fines
 - Imprisonment
 - Loss of grant funding
 - Loss of access to data
 - Loss of employment
 - Loss of reputation
- Institutional consequences
 - Loss of institutional access to data
 - Department/school faculty
 - Entire university faculty
 - Loss of grant funding
 - Loss of federal or state financial support for research
 - Loss of confidence in university research

Restricted Data Use History

Tracking researchers’ prior restricted data use is something a handful of data repositories are currently doing, but only internally. Three repositories require researchers to state explicitly in their data access requests that they have experience with restricted data, and one requires evidence of that usage through publication citations based on those data. Incidentally, none of these three repositories have a requirement that researchers complete or certify completion of any data security or restricted data handling training. For repositories where the same researchers request multiple datasets over their careers, the repository staff who evaluate the data access requests do have their own internal tracking and familiarity with whether those researchers properly used the restricted data, but this information is not shared with other institutions and it is generally not required of the researchers excepting the three previously mentioned institutions. Despite the lack of requirements for tracking and sharing this information, repository representatives in the interviews and at the July 2017 Georgetown University convening highlighted a record of good data use, including which datasets have been

used and any “bad data handling” (disclosive events and non-disclosive acts), as an important component of any transferrable researcher credential.

Perspectives on importance of prior data use history

- *“I guess probably you would feel like they've already successfully handled restricted data and if there's been no breach or no problems that they have experience that the others don't have. ... I think in part it might be just even just the confidentiality of the data, you can let somebody maybe, you feel more comfortable with them getting certain data sets that have less risky stuff in it when their experience is less.” (SF_008_02)*
- *“Yeah, history of prior data use and documentation of training, particularly around hygiene with regard to using restricted data, I think would be very useful.” (SF_007_01)*
- *“Which does not necessarily mean that it didn't happen, but basically that would also be a breach of the contract. If we get to know it then it might mean that one of our penalties applies. Usually, in that case, the penalties that are really hard, are usually that you won't get access for any project in the next two years or something like that. That, of course, can be very hurtful, also that your institution cannot get restricted access, which also means that other researchers from that institution cannot access the data, if we have the feeling that this is not just a minor breach but a major one. What we also potentially do, is tell other research data centers about the incident, which can also be kind of not so nice, if everybody knows that you've been working with the data but not been very ... What do you say ... If everybody in Germany for example knows this, that can be hard for your reputation.” (SF004_1)*

Solutions

Language Harmonization

Our analysis of repository policies found that the language used to describe levels of data restriction, confidentiality, and access methods differs significantly both among and within repositories. This language inconsistency confounds the challenge of developing a transferable digital researcher identity. A critical next step is standardizing the terminology used to describe the elements of restricted data security and access. Establishing a common set of terms and definitions, which will necessarily evolve over time through collaboration within the research community, will allow different repositories to understand and integrate shared standards and technologies into their own processes.

We recommend the following language to designate *levels of risk*, within which individual repositories can fit their requirements for data security and data protection. These levels reflect a combination of the sensitivity of the data, the probability of re-identification, and legal mandates.

- Low
 - Data with no identifiable information included. The risk of re-identification, disclosure, and/or harm to research subjects is minimal, and no legal or statutory limitations apply. These data are generally available for public access.
- Moderate
 - Data that have not been fully anonymized and the risk of re-identification, disclosure, and/or risk to research subjects is not trivial. These data may require a data access request, license, data use agreement, or other formal process.
- High
 - Data that do not include direct identifiers (e.g., name, social security number) but that have not been otherwise anonymized, and so pose a higher risk of re-identification, disclosure, and/or harm to research subjects. These data require in-depth review of project materials, approval by data provider and/or data repository, signed legal agreements between the user and the repository, and perhaps government-issued clearances.
- Highest
 - Data that include direct identifiers, data that could be relatively easily re-identified *and* cover sensitive subjects whose disclosure could harm research subject, or data that are governed by very restrictive legal requirements for third-party access. These data require special, secure handling due to one or

more factors. Access is limited to secure computing facilities and physical enclaves, in-depth identity verification, and careful vetting of output or other materials removed from the secure environment.

Shared Understanding of Data Security

To the extent that a dataset's analytical risk can be classified, we encourage repositories and research analysts to leverage standardized tools in their work. For example, once available, Harvard's [DataTags](#) tools, described in this paper's introduction, will allow repositories to apply an objective risk assessment tool to their management of data protection and accessibility. See Appendix 2 for a simple mapping of ICPSR's security levels to those used in DataTags and in New York University Center for Urban Science + Progress's Data Governance and Confidentiality Policy.⁹ For research analysts, guidance on the use of a 'privacy budget' predicated on the mathematical concept of differential privacy (described in this paper's fourth footnote) will allow them to more precisely balance, and in fact to optimize, both the privacy of research subjects and the need for robust research in the public interest.

A Durable and Transferable Digital Identifier

A number of researcher attributes emerged as common elements of user access requirements across repositories. We recommend the creation and broad use of a durable and transferable digital researcher identifier. This will require verification of identity, education, and professional affiliation, along with capture of publication citations, evidence of data use, recognized trainings completed, and other elements used to establish the researcher's reputation as a trusted data steward. These attributes, already commonly captured by individual repositories, should be verified, stored, and kept up-to-date by a secure, central system accessible by participating researchers and data stewards.

Professional Roles, Demonstrated Trust, and Track Record of Responsible Use

The basic elements of the researcher ID are individual researcher attributes such as citizenship/visa status, academic degree(s), institutional affiliation(s), federal or private grants awarded, indicators of demonstrated responsible use of restricted data, and perhaps prior data sharing and other contributions to the research community. When linked to existing researcher IDs including ICPSR's MyData system and the ORCID system, it will also seamlessly incorporate publications and other citations.

⁹ https://datahub.cusp.nyu.edu/sites/default/files/documents/policies/Data_Governance.pdf

Training Requirements

Shared training standards relevant across datasets and access environments will be a key element of a trusted researcher identity and must be identified or developed in collaboration with recognized stakeholders and experts. The issuance of visas for particular datasets may entail additional, data-specific training. Training programs that meet community-normed standards will provide critical instruction for the protection of data and will encourage the development of a **culture of confidentiality**. Risk and magnitude of potential harm to studied entities are difficult to quantify, requiring researchers and their institutions to internalize the value of protecting confidentiality and to hold themselves and their peers to a high standard of care. This standard of care is demonstrated in most elements of the research endeavor, and training researchers to be excellent data stewards should be no exception. At present, where there are training requirements, they focus on an understanding of the relationship of privacy and confidentiality, research subject risk assessment, and best practices for physical and network-based data security. These topics can form the baseline for restricted data user training if users consistently encounter them. The content should be expanded to meet the goal of establishing a user culture of confidentiality and data stewardship.

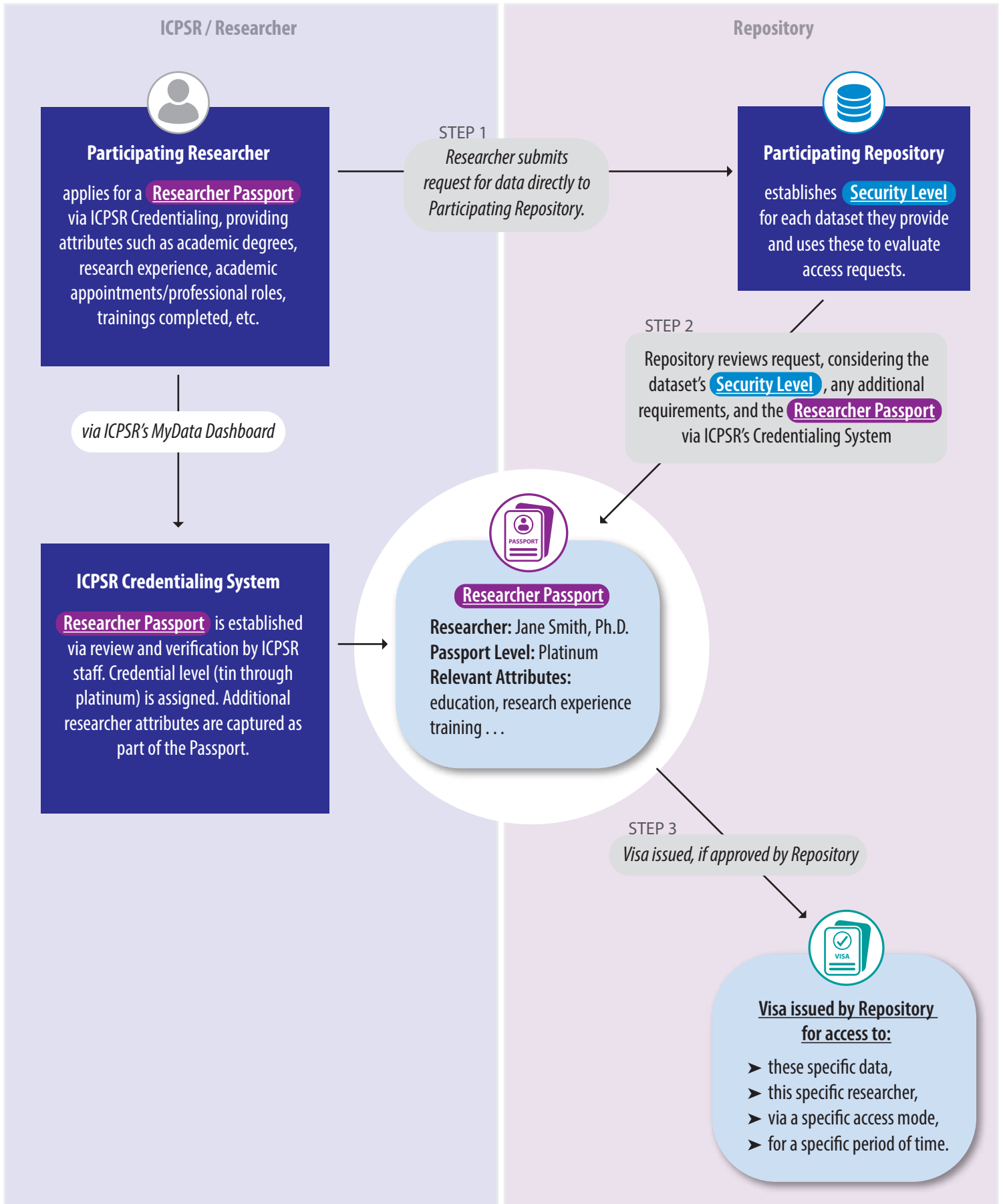
Overview of the Researcher Passport and Visa Processes

The Researcher Passport indicates both overall level of “trust” and specific characteristics that may affect terms of access to particular datasets. The passport level (tin, bronze, silver, gold, or platinum) reflects the level of trust achieved by the researcher and status with regard to those characteristics.

The Visa is issued by the data custodian in accordance with their assessment of the sensitivity and legal and other governance of the dataset. Recommended modes of access to data of different risk levels for researchers with different passports are provided, including guidance for the mode of access for both the Principal Investigator (or instructor, for classroom use) and for the research team (or students).

The following pages describe the workflow of the Researcher Passport and Visa system. See Appendix 3 for a set of six use cases illustrating practical application of the system.

The Researcher Passport and Visa Provisioning System: A Roadmap





The Researcher Passport

Trustworthiness of Individual *Question asked: How trustworthy is the researcher?*

The Researcher Passport sets forth a set of common characteristics of data users to be recorded and verified for each credentialed user. Together, they aim to represent a particular user's history of research experience, data stewardship, and education and training that has over time established his/her trustworthiness as a research professional. The research underpinning the project described in this white paper illuminates shared goals for provisioning data access to researchers, and this passport rubric aims to standardize the way data users build, and data repositories assess, trust.

USER ATTRIBUTES	POINTS ATTRIBUTED
Highest degree earned	
Doctoral/terminal degree	3
Graduate degree (non-terminal)	2
Undergraduate	1
No degree	0
Professional Position (choose one of the following two options)*	
Option 1 Academic faculty/staff: Highest institutional appointment/affiliation	
Full/Associate professor	3
Assistant professor	2
Student	1
Research staff	1
Option 2 Non-profit, for-profit, government, or media staff: Years of relevant experience	
5+	3
3-4	2
0-2	1
Other	
Recognized Federal clearances	4
Current (2 pts) or recent (1 pt) Federal grant	2/1
Research publications (1 or more publications)	2
Restricted data use experience (1 or more projects)	2
Potential dataset- or repository-specific user requirements	
Country- or region-specific citizenship or residency status	specify
Affiliation with Carnegie-classified academic institution	yes/no
Badges earned and verified	
Trainings	
Data security — Levels I-III	specify
Research conduct — Levels I-III	specify
Other	specify
Specific expertise	
Restricted qualitative data use	specify
Other	specify
Contributions — data stewardship	
History of data sharing	citation/DOI
History of metadata enhancement	citation/DOI
History of code/syntax sharing	citation/DOI
Confirmed research misconduct (unintentional procedural violations and/or intentional data disclosure or misuse)	yes/no



Passport status and issuance based upon the points earned in the credentialing process

*Users should select the institutional realm (academic appointment v. non-profit, etc. years of experience) that best represents the institution under which the majority of their research will be conducted.



Establishing Security Level of the Data (by Repository)

Question asked: What security level do the data require?

Levels of required security measures are a key factor in the decision-making process for the provisioning of restricted data. This scoring system aims to guide repositories in the assessment and ranking of their datasets. This will facilitate alignment of credentialed users with the data they request, by recommending whether to provide access and if so, the most appropriate mode of access given the combination of the two.

DATA CHARACTERISTICS	POINTS ATTRIBUTED
Sensitivity level	If yes, then add...
protected population	+ 3
proprietary data	+ 4 to 6
potentially harmful personal information	+ 4
Disclosure risk level	
sample size	+ 1 to 4
geographic region size	+ 1 to 4
rare sample attributes	+ 1 to 4
link to public data	+ 3
Legal or statutory limitations	
HIPAA	+ 6
FERPA	+ 6
other legislated restrictions	+ 3 to 6
.....	
Data security score (after totalling above)	Range
low	0-3
moderate	4-5
high	6-9
highest	10+



Visa Application & Issuance

Questions asked: Based on the established level of trustworthiness, is it recommended that the repository issue a Visa? If yes, what access modality is recommended based on both the trustworthiness of the user and the security level of the data?

The following tables represent standard recommendations made to repositories for access modalities based on the relationship between the user passport level and the security level of the requested data. That relationship then informs a recommendation for the most appropriate mode of access.



Visas issued by data custodians in accordance with the user's credential and the data security level and requirements

It is important to note that this matrix is not intended to replace repository review of access requests but to provide a conceptual model to facilitate their review. Repositories are offered this standardized and community-normed system to inform their decision making on whether to issue a Visa: access to a specific dataset, to a specific researcher, in a specific environment, and for a specific amount of time.

Individual PI's Access

Principal Investigator (PI) Access Matrix

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin



Research Team Member (RTM) Access Matrix

RTM SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	virtual enclave*	no access	Bronze
4	unrestricted	virtual enclave	virtual enclave*	no access	Copper
0-3	unrestricted	virtual enclave*	virtual enclave*	no access	Tin

*Access to Moderate/High Security level data in a virtual environment if PI holds required credential

Potential additional data-specific considerations:

- Does this dataset require the user to be a citizen of a certain country?
- Does this dataset require a certain level or kind of training?
- Does this dataset require a certain Federal Clearance?
- Does this dataset require a certain mode of access?

Building the System

ICPSR will design, build, and maintain the underlying technology to support a credentialing system within the existing MyData¹⁰ system. This digital identification system is required for users to access ICPSR data, whether public or restricted. In the past, it has been used primarily to control access to ICPSR's membership-only archive. MyData maintains the user's history of data searches and downloads, allowing ICPSR to proactively provide users with study updates and information about new data, related citations, and other relevant opportunities. Once the credentialing system is implemented, MyData account holders who meet required criteria will be designated with the appropriate digital passport, allowing them to demonstrate and attest to various elements of their identities and allowing data repositories and other providers to validate the identities and credentials of individual researchers. This system will also allow for individual visas to be issued once an investigator has applied for and been granted access to a certain dataset.

ICPSR's updated and expanded MyData system will serve as the researcher passport home. Through this system's dashboard, researchers will input relevant identifying and supporting information (name, institution, appointment, degrees, citizenship status, relevant security clearances obtained, etc.). The passport will detail training programs along with dates completed, federal grants received, records of current or prior restricted data access, in addition to the researchers' publications, archived datasets, shared analysis files or code, etc. Incorporation of a researcher's ORCID¹¹ identifier will strengthen the passport's ability to find and import citations beyond what ICPSR already captures. All information submitted by the researcher as part of the passport application will be held securely and made available only to third parties (e.g., other data custodians) who are participating in the researcher credentialing system.

Each researcher accesses his or her own dashboard through the MyData account. ICPSR will build a similar interface through which participating repositories can view a researcher's passport. There they will find all publicly-facing, relevant elements of the researcher's identity that will facilitate efficient and accurate understanding of his/her qualifications. The elements of this credential will allow ICPSR to locate them within a matrix, with the y axis representing increasing community-normed levels of researcher experience and training, and the x axis representing increasing community-normed levels of data disclosure risk and necessary protections (see Principle Investigator Access Matrix, page 23). The repository can then use this shared understanding to evaluate a researcher's fitness for access to its data and identify the appropriate method of access. Each repository will define its own requirements for issuing visas, which may vary by dataset, but must include the following information which will be transmitted electronically to the passport database: datasets

¹⁰ ICPSR MyData Account: <https://www.icpsr.umich.edu/cgi-bin/newacct>

¹¹ ORCID: <https://orcid.org/>

provided, beginning and end dates of access, access modality, output review process, and project scope. Once a project is complete, the repository will transmit any records of data use agreement or security plan violations.

An important element of the researcher passport will be a broadly accepted system for evaluating allegations of research misconduct (both unintentional procedural violations and intentional data disclosure or misuse) and allowing for due process if allegations are disputed. The system will be implemented by a committee with membership representative of the consortium of participating institutions and the passport will reflect any confirmed violations for a predetermined period of time. Consequences of research misconduct are often legally assumed only by the institution with which the researcher is affiliated. Associating misconduct with the researcher's reputation via the passport is intended to encourage further responsible data stewardship by the investigator and research team members, reinforcing any legal agreements in place, as well as alerting other repositories to verified misconduct.

A mechanism for updating passports will be established. The researcher will update information on degrees earned, changes in institutional affiliations or appointments, trainings completed, or citizenship status. Data usage history and visas will be updated by the repository, and changes in level (e.g., gold to platinum) or where the researcher falls within the aforementioned matrix, will be determined by the passport issuer, triggered either by updates made to the passport or by a direct request from the researcher.

Benefits

The primary benefit to addressing these three challenges – reconciling how repositories classify restricted data, developing a method to communicate researcher reputation, and standardizing training requirements – is that in doing so, participating repositories will have a basis on which to trust the credentials created and verified by other institutions. The existing data access request process has too many conflicting and confusing policies and procedures. These limit interoperability among repositories. A researcher who has been approved at one repository to access “restricted” data must repeatedly verify this information to each repository, there is no guarantee that “restricted” means the same thing at the new institution, the new institution has limited ability to verify that there have not been incidences of poor data stewardship, and the training requirements may differ.

Standardizing these three requirements will streamline the process for both the researcher and the repository staff. Training can be more complete, directly relevant, and less bureaucratic. Access procedures can be more transparent when repositories use harmonized language and processes.

Standardizing the information required from researchers and the validation process will enable repository staff to streamline the review process. The process now includes two phases: verify and validate all the portions of the data access request that identify the researchers, and verify and validate the remaining components that reflect the specific research project proposal. By standardizing terminology, training requirements, and how researcher data use histories are recorded and shared, repositories can create standardized review processes that make it easier and more reasonable to accept other repositories’ credentials because the verification and validation processes are the same.

Next Steps for the Credentialing Project

New Restricted Data Use Agreement Template

Common Data Use Agreements (DUAs) will streamline access to restricted data. To that end, ICPSR has adopted a template DUA, included here in Appendix 4. It can be modified for use by other repositories.

Training Evaluation and Certification

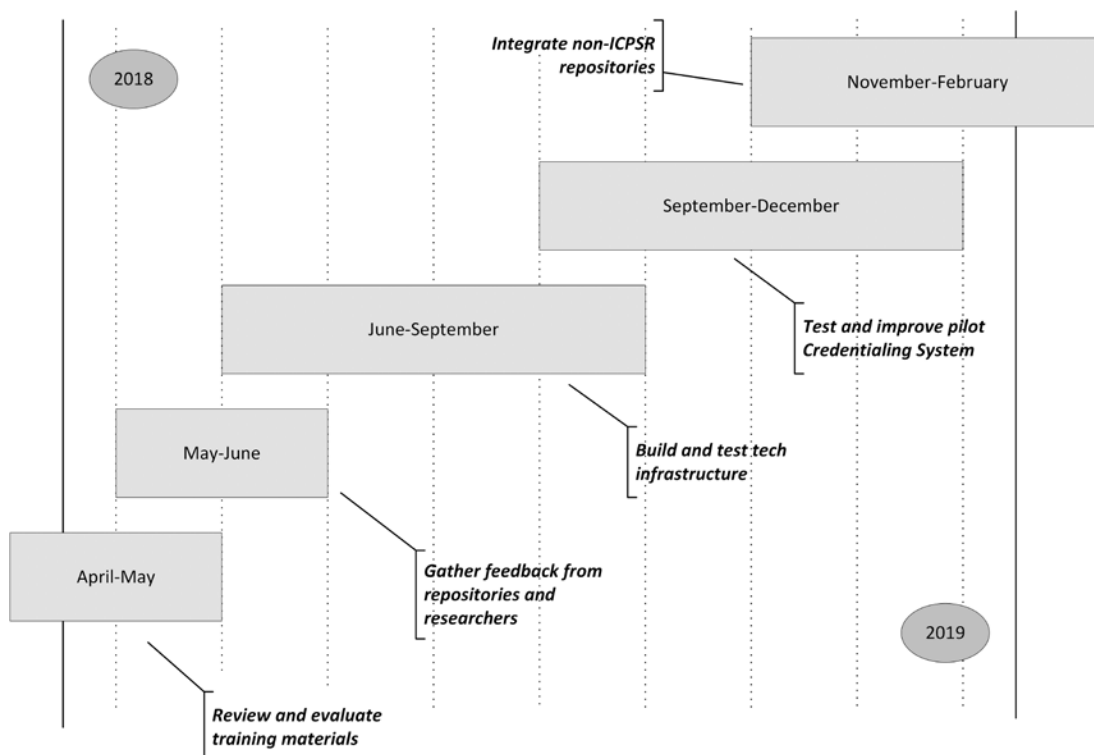
Review and evaluation of training materials (PEERRS, CITI, and repository-created training materials) are underway in order to define the minimum requirements for training on data security and supporting a culture of confidentiality and data stewardship. This review will result in more explicit requirements and guidelines for the completion and sharing of training modules between repositories through the credentialing system.

Outreach and Testing

ICPSR will work closely with other repositories to pilot the credentialing system. Enhancements and improvements will be made to the technical infrastructure, workflow, and governing policies and procedures.

Timeline

FIGURE 2. IMPLEMENTATION TIMELINE

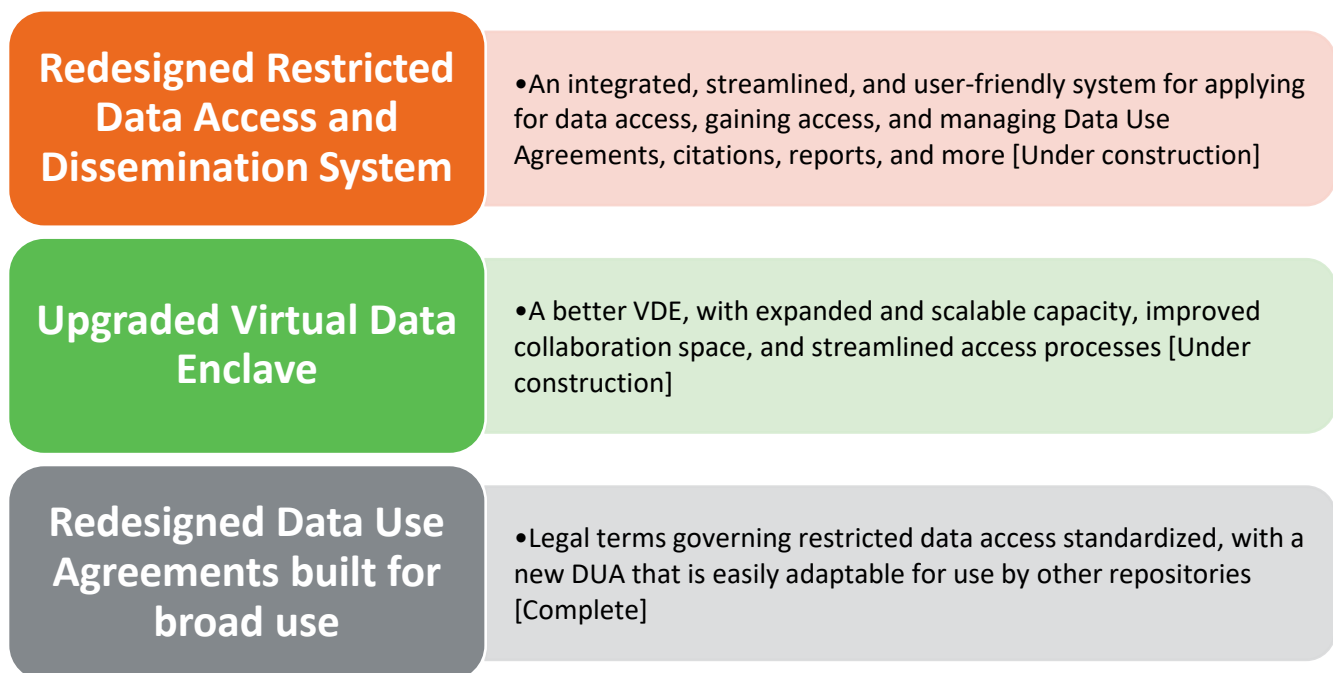


Conclusion

The creation of a broadly accepted researcher credential aims to ease the burden and remove many of the roadblocks associated with responsible and efficient use of secondary data for the social science research endeavor. The objective of this project is to encourage and foster a culture of confidentiality and shared responsibility for good data stewardship through community-normed terminology and the design and implementation of a broadly adopted researcher passport system.

The social science research community is facing unprecedented growth and development of data types and sources, as well as researchers with creative and complex plans for their use. As a result, there are equally complex challenges to confidentiality protection, data security, and research integrity. As the United States government, along with other governments and data repositories around the world, call both for increased data sharing and for more sophisticated means of protecting privacy, the community can choose to respond collaboratively to ease the increasing burden on researchers and data repositories. Shared language, policies, and tools for responsibly and securely providing appropriate data access to social scientists will help eliminate the redundancies in processes and recover scarce resources that are better applied to the pursuit of science than to antiquated and cumbersome procedural hoops.

FIGURE 3. NEW ICPSR TOOLS AND SYSTEMS, IMPROVING RESTRICTED DATA ACCESS



Appendix 1

Data Collection Sites

- Inter-university Consortium for Political & Social Research (ICPSR)
- University of Michigan's Institute for Research on Innovation & Science
- Michigan Center on the Demography of Aging
- UK Data Archive
- Federal Statistical Research Data Centers
- Cornell Institute for Social & Economic Research
- Minnesota Population Center
- DataVerse
- The Demographic & Health Surveys Program
- NORC at the University of Chicago
- Australian National Data Service
- GESIS – Leibniz Institute for the Social Sciences
- Statistische Ämter des Bundes und der Länder Forschungsdatenzentren (German Federal Statistical Offices)
- Databases of Genotypes & Phenotypes
- Protein Data Bank
- Research Data Assistance Center
- Deep Blue (University of Michigan Libraries)
- Cambridge (UK) Data Archive
- New York University's Center for Urban Science and Progress (CUSP)
- Australian Data Archive
- Substance Abuse and Mental Health Data Archive/National Survey on Drug Use and Health
- National Data Archive on Child Abuse and Neglect

Appendix 2

NYU CUSP	ICPSR Passports/Access	Harvard DataTags
<p>Green/Restricted Green</p> <p>NYC OpenData; city agency data w/ no identifiers & nonsensitive information but not NYC OpenData; city agency data w/ no identifiers & nonsensitive information with add'l access requirements specified in agreement; public data from other sources containing non-personal information.</p>	<p>Low</p> <p>Unrestricted</p>	<p>Blue</p> <p>Non-confidential information, stored and shared freely.</p> <p>Green</p> <p>Not harmful personal information, shared with some access control.</p>
<p>Yellow</p> <p>De-identified education microdata; de-identified health records; disclosable aggregate research datasets; disclosable reporting databases for APIs/webtools; city agency data which are de-identified but contain sensitive personal information; audio recordings or images containing individual information.</p>	<p>Moderate</p> <p>Provided via secure download to Silver through Platinum researchers; provided via virtual enclave to Copper and Bronze researchers.</p>	<p>Yellow</p> <p>Potentially harmful personal information, shared with loosely verified and/or approved recipients.</p>
<p>Red</p> <p>Microdata containing personal information or any other direct identifiers; personally identifiable information data.</p>	<p>Highest</p> <p>Provided via physical enclave to Silver researchers; provided via virtual or physical enclave to Gold and Platinum researchers.</p>	<p>Orange</p> <p>Sensitive personal information, shared with verified and/or approved recipients under agreement.</p>
		<p>Red</p> <p>Very sensitive personal information, shared with strong verification of approved recipients under signed agreement.</p> <p>Crimson</p> <p>Maximum sensitive, explicit permission for each transaction, strong verification of approved recipients under signed agreement.</p>

Appendix 3

The Researcher Passport and Visa: Use Cases

Six use cases ranging from a well-established academic researcher to a media-based analyst using the National Intimate Partner & Sexual Violence Survey (NISVS), a moderate security dataset.

NISVS Use Case 1 — Faculty at the University of Pittsburgh

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	Ph.D.	3
Professional position	Associate Prof.	3
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	yes - recent	1
Research publications (1 or more publications)	yes	2
Restricted data use experience (1 or more projects)	none	0
Potential dataset- or repository-specific user requirements		total = 9 pts
Country- or region-specific citizenship or residency status	US citizen	
Affiliation with Carnegie-classified academic institution	yes	
Badges earned and verified		
Trainings*		
Data security — Levels I-III	level I	
Research conduct — Levels I-III	level II	
Specific expertise		
Contributions — data stewardship		
History of data sharing	DOI	
History of metadata enhancement	no	
History of code/syntax sharing	DOI	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	do not know	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin

*Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.

VISA RECOMMENDATION

User	Total points	9
	User's Passport level	Platinum
Data	Security level of NISVS data	Moderate
Visa	Visa recommended for this user + these data?	Yes
	Mode of access recommended for this user/data combination?	Secure Download

NISVS Use Case 2 — Pre-candidacy Doctoral Student at Arizona State University

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	B.A.	1
For-Profit: Years of relevant experience	Student	1
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	none	0
Research publications (1 or more publications)	yes	2
Restricted data use experience (1 or more projects)	none	0
Potential dataset- or repository-specific user requirements		total = 4 pts
Country- or region-specific citizenship or residency status	US citizen	
Affiliation with Carnegie-classified academic institution	yes	
Badges earned and verified		
Trainings*		
Data security — Levels I-III	level I	
Research conduct — Levels I-III	level II	
Specific expertise		
Contributions — data stewardship		
History of data sharing	no	
History of metadata enhancement	no	
History of code/syntax sharing	no	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	★	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin

*Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.

VISA RECOMMENDATION

User	Total points	4
	User's Passport level	Copper
Data	Security level of NISVS data	Moderate
Visa	Visa recommended for this user + these data?	Yes
	Mode of access recommended for this user/data combination?	VDE with credentialed PI

NISVS Use Case 3 — Staff Researcher at For-profit Research Institution

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	Ph.D.	3
For-Profit: Years of relevant experience	5 yrs	3
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	yes - current	2
Research publications (1 or more publications)	yes	2
Restricted data use experience (1 or more projects)	yes	2
Potential dataset- or repository-specific user requirements		total = 12 pts
Country- or region-specific citizenship or residency status	Chinese citizen, currently living in US (1 yr)	
Affiliation with Carnegie-classified academic institution	no	
Badges earned and verified		
Trainings*		
Data security — Levels I-III	level III	
Research conduct — Levels I-III	level II	
Specific expertise	n/a	
Contributions — data stewardship		
History of data sharing	DOI	
History of metadata enhancement	no	
History of code/syntax sharing	no	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin

*Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.

VISA RECOMMENDATION

User	Total points	12
	User's Passport level	Platinum
Data	Security level of NISVS data	Moderate
Visa	Visa recommended for this user + these data?	Yes
	Mode of access recommended for this user/data combination?	Secure Download

NISVS Use Case 4 — Community College Professor (data access for class learning)

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	M.A.	2
Professional position	Asst. Prof.	2
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	none	0
Research publications (1 or more publications)	yes	2
Restricted data use experience (1 or more projects)	none	0
Potential dataset- or repository-specific user requirements		total = 6 pts
Country- or region-specific citizenship or residency status	US citizen	
Affiliation with Carnegie-classified academic institution	no	
Badges earned and verified		
Trainings		
Data security — Levels I-III	level I	
Research conduct — Levels I-III	level II	
Specific expertise		
Contributions — data stewardship		
History of data sharing	no	
History of metadata enhancement	no	
History of code/syntax sharing	no	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	no access	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin


**Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.*

VISA RECOMMENDATION

User	Total points	6
	User's Passport level	Silver
Data	Security level of NISVS data	Moderate
Visa	Visa recommended for this user + these data?	Yes
	Mode of access recommended for this user/data combination?	Secure download
Student Access	Mode of access recommended for this user's research team?	VDE

NISVS Use Case 5 — Graduate Student at Beijing University

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	M.A.	2
Professional position	Grad Student	1
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	none	0
Research publications (1 or more publications)	none	0
Restricted data use experience (1 or more projects)	none	0
Potential dataset- or repository-specific user requirements		total = 3 pts
Country- or region-specific citizenship or residency status	Chinese citizen	
Affiliation with Carnegie-classified academic institution	no	
Badges earned and verified		
Trainings		
Data security — Levels I-III	level I	
Research conduct — Levels I-III	level II	
Specific expertise	n/a	
Contributions — data stewardship		
History of data sharing	no	
History of metadata enhancement	no	
History of code/syntax sharing	no	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted	virtual enclave	no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted		no access	no access	Tin

**Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.*

VISA RECOMMENDATION

User	Total points	3
	User's Passport level	Tin
Data	Security level of NISVS data	Moderate
	Visa recommended for this user + these data?	No
Visa	Mode of access recommended for this user/data combination?	n/a

NISVS Use Case 6 — Staff Reporter at the *New York Times*

USER ATTRIBUTES	RESULTS	POINTS
Highest degree held	B.A.	1
Professional position: Media years experience	3	2
Other		
Recognized Federal clearances	none	0
Current (2pts) or recent (1pt) Federal grant	none	0
Research publications (1 or more publications)	none	0
Restricted data use experience (1 or more projects)	yes	2
Potential dataset- or repository-specific user requirements		total = 5 pts
Country- or region-specific citizenship or residency status	Spanish citizen, currently living in US (3 yrs)	
Affiliation with Carnegie-classified academic institution	no	
Badges earned and verified		
Trainings		
Data security — Levels I-III	level I	
Research conduct — Levels I-III	level I	
Specific expertise	n/a	
Contributions — data stewardship		
History of data sharing	no	
History of metadata enhancement	no	
History of code/syntax sharing	no	
Confirmed research misconduct	no	

PI SCORE	DATA SECURITY LEVEL				
	LOW	MODERATE	HIGH	HIGHEST	
8+	unrestricted	secure download	secure download	VDE / physical enclave	Platinum
7	unrestricted	secure download	virtual enclave	physical enclave	Gold
6	unrestricted	secure download	virtual enclave	no access	Silver
5	unrestricted		no access	no access	Bronze
4	unrestricted	virtual enclave	no access	no access	Copper
0-3	unrestricted	no access	no access	no access	Tin

**Researching and recommending training standards, contents, and levels is part of year two of this project. The levels listed here act as placeholders until that work is complete.*

VISA RECOMMENDATION

User	Total points	5
	User's Passport level	Bronze
Data	Security level of NISVS data	Moderate
Visa	Visa recommended for this user + these data?	Yes
	Mode of access recommended for this user/data combination?	VDE

Appendix 4

Restricted Data Use Agreement for Restricted Data from <REPOSITORY NAME>

I. Definitions

A. “Investigator” is the person primarily responsible for conducting the research or statistical activities relative to the Research Description of the <APPLICATION NAME> (the “Research Description”), or supervising the individuals conducting the research or statistical activities relative to the Research Description, for which Restricted Data are obtained through this Agreement.

B. “Research Staff” are all persons at the Investigator's Institution, excluding the Investigator, who will have access to Restricted Data obtained through this Agreement, including students, other faculty and researchers, staff, agents, or employees for which Institution accepts responsibility.

C. “Institution” is the university or research institution at which the Investigator will conduct research using Restricted Data obtained through this Agreement.

D. “Representative of the Institution” is a person authorized to enter into binding legal agreements on behalf of Investigator's Institution.

E. “Restricted Data” are the research dataset(s) provided under this Agreement that include potentially identifiable information in the form of indirect identifiers that if used together within the dataset(s) or linked to other dataset(s) could lead to the re-identification of a specific Private Person, as well as information provided by a Private Person under the expectation that the information would be kept confidential and would not lead to harm to the Private Person. Restricted Data includes any Derivatives.

F. “Private Person” means any individual (including an individual acting in an official capacity) and any private (i.e., non-government) partnership, corporation, association, organization, community, tribe, sovereign nation, or entity (or any combination thereof), including family, household, school, neighborhood, health service, or institution from which the Restricted Data arise or were derived, or which are related to a Private Person from which the Confidential Information arise or were derived.

G. “<REPOSITORY ACRONYM>” is <REPOSITORY NAME>.

H. “<APPLICATION NAME>” includes all information entered into the <REPOSITORY ACRONYM> data access request system, including Investigator information, Research Staff information, Research Description, Data Selection specifying which files and documentation are requested, Confidentiality Pledge signed by the Investigator, Supplemental Agreement and Confidentiality Pledge signed by each Research Staff, Data Security Plan, and a copy of a

document signed by the Institution's Institutional Review Board (IRB), or equivalent, approving or exempting the research project.

I. "Data Security Plan" is a component of the Agreement which specifies permissible computer configurations for use of Restricted Data and records what the Investigator commits to do in order to keep Restricted Data secure.

J. "Deductive Disclosure" is the discerning of a Private Person's identity or confidential information through the use of characteristics about that Private Person in the Restricted Data. Disclosure risk is present if an unacceptably narrow estimation of a Private Person's confidential information is possible or if determining the exact attributes of the Private Person is possible with a high level of confidence.

K. "Derivative" is a file or statistic derived from the Restricted Data that poses disclosure risk to any Private Person in the Restricted Data obtained through this Agreement. Derivatives include copies of the Restricted Data received from <REPOSITORY ACRONYM>, subsets of the Restricted Data, and analysis results that do not conform to the guidelines in Section VI.F.

II. Responsibility to Address Disclosure Risk

Deductive Disclosure of a Private Person's identity from research data is a major concern of federal agencies, researchers, and Institutional Review Boards. Investigators and Institutions who receive any portion of Restricted Data are obligated to protect the Restricted Data from Deductive Disclosure risk, non-authorized use, and attempts to identify any Private Person by strictly adhering to the obligations set forth in this Agreement.

III. Requirements of Investigator

- A. The Investigator assumes the responsibility of completing the <APPLICATION NAME> and any other required documents, reports, and amendments.
- B. The Investigator agrees to manage and use Restricted Data, implement all Restricted Data security procedures per the Data Security Plan, and ensure that all Research Staff understand their requirements per this Agreement and follow the Data Security Plan.
- C. Investigators must meet each of the following criteria:
 - 1. Have a PhD or other research-appropriate terminal degree; and
 - 2. Hold a faculty appointment or have an appointment that is eligible to be a principal investigator at Institution.

IV. Requirements of Institution

The Institution represents that it is:

- A. An institution of higher education, a research organization, a research arm of a government agency, or a nongovernmental, not-for-profit, agency.
- B. Not currently debarred or otherwise restricted in any manner from receiving information of a sensitive, confidential, or private nature under any applicable laws, regulations, or policies.
- C. Have a demonstrated record of using sensitive data according to commonly accepted standards of research ethics and applicable statutory requirements.

V. Obligations of <REPOSITORY ACRONYM>

In consideration of the promises made in Section VI of this Agreement, and upon receipt of a complete and approved <ONLINE APPLICATION>, <REPOSITORY ACRONYM> agrees to:

- A. Provide the Restricted Data requested by the Investigator in the Restricted Data Order Summary within a reasonable time of execution of this Agreement by Institution and to make the Restricted Data available to Investigator via download or removable media.
- B. Provide electronic documentation of the origins, form, and general content of the Restricted Data sent to the Investigator, in the same time period and manner as the Restricted Data.

<REPOSITORY ACRONYM> MAKES NO REPRESENTATIONS NOR EXTENDS ANY WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE RESTRICTED DATA WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS. Unless prohibited by law, Institution assumes all liability for claims for damages against them by third parties that may arise from the use, storage, disposal, or disclosure by the Institution of the Restricted Data, except to the extent and in proportion such liability or damages arise from the negligence of <REPOSITORY ACRONYM>.

VI. Obligations of the Investigator, Research Staff, and Institution

Restricted Data provided under this Agreement shall be held by the Investigator, Research Staff, and Institution in strictest confidence and can be used or disclosed only in compliance with the terms of this Agreement. In consideration of the promises in Section V of this Agreement, and for use of Restricted Data from <REPOSITORY ACRONYM>, the Institution agrees:

- A. That the Restricted Data will be used solely for research or statistical purposes relative to the project as identified in the Research Description of the <ONLINE APPLICATION> (the “Research Description”), and for no other purpose whatsoever without the prior written consent of <REPOSITORY ACRONYM>. Further, no attempt will be made to identify Private Person(s), no Restricted Data of Private Person(s) will be published or otherwise distributed, the Restricted Data will be protected against Deductive Disclosure risk by strictly adhering to the obligations set forth in this Agreement, and precautions will be taken to protect the Restricted Data from non-authorized use.

- B. To comply fully with the approved Data Security Plan at all times relevant to this Agreement.
- C. That no persons other than those identified in this Agreement or in subsequent amendments to this Agreement, as Investigator or Research Staff and who have signed this Agreement or a Supplemental Agreement, be permitted access to the contents of Restricted Data files or any Derivatives from the Restricted Data.
- D. That within five (5) business days of becoming aware of any unauthorized access, use, or disclosure of Restricted Data, or access, use, or disclosure of Restricted Data that is inconsistent with the terms and conditions of this Agreement, the unauthorized or inconsistent access, use, or disclosure of Restricted Data will be reported in writing to <REPOSITORY ACRONYM>.
- E. That, unless prior specific, written approval is received from <REPOSITORY ACRONYM>, no attempt under any circumstances will be made to link the Restricted Data to any Private Person, whether living or deceased, or with any other dataset, including other datasets provided by <REPOSITORY ACRONYM>.
- F. To avoid inadvertent disclosure of Private Persons by being knowledgeable about what factors constitute disclosure risk and by using disclosure risk guidelines, such as but not limited to, the following guidelines¹ in the release of statistics or other content derived from the Restricted Data.²
1. No release of a sample unique for which only one record in the Restricted Data provides a certain combination of values from key variables.
 2. No release of a sample rare for which only a small number of records (e.g., 3, 5, or 10 depending on sample characteristics) in the Restricted Data provide a certain combination of values from key variables. For example, in no instance should the cell frequency of a cross-tabulation, a total for a row or column of a cross-tabulation, or a quantity figure be fewer than the appropriate threshold as determined from the sample characteristics. In general, assess empty cells and full cells for disclosure risk stemming from sampled records of a defined group reporting the same characteristics.
 3. No release of the statistic if the total, mean, or average is based on fewer cases than the appropriate threshold as determined from the sample characteristics.
 4. No release of the statistic if the contribution of a few observations dominates the estimate of a particular cell. For example, in no instance should the quantity figures be released if one case contributes more than 60 percent of the quantity amount.
 5. No release of data that permits disclosure when used in combination with other known data. For example, unique values or counts below the appropriate threshold for key variables in

¹ For more information, see the U.S. Bureau of the Census checklist. *Supporting Document Checklist on Disclosure Potential of Data*, at http://www.census.gov/srd/sdc/S14-1_v1.3_Checklist.doc; *NCHS Disclosure Potential Checklist* at http://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02A.pdf; and *FCSM Statistical Policy Working Paper 22 (Second Version, 2005)* at <http://www.hhs.gov/sites/default/files/spwp22.pdf>

² If disclosure review rules were established for a specific Restricted Dataset, they will be included in the dataset's documentation and are covered by this Agreement.

the Restricted Data that are continuous and link to other data from <REPOSITORY ACRONYM> or elsewhere.

6. No release of minimum and maximum values of identifiable characteristics (e.g., income, age, household size, etc.) or reporting of values in the “tails,” e.g., the 5th or 95th percentile, from a variable(s) representing highly skewed populations.
 7. No release of ANOVAs and regression equations when the analytic model that includes categorical covariates is saturated or nearly saturated. In general, variables in analytic models should conform to disclosure rules for descriptive statistics (e.g., see #6 above).
 8. In no instance should data on an identifiable case, or any of the kinds of data listed in preceding items 1-7, be derivable through subtraction or other calculation from the combination of tables released.
 9. No release of sample population information or characteristics in greater detail than released or published by the researchers who collected the Restricted Data. This includes but is not limited to publication of maps.
 10. No release of anecdotal information about a specific Private Person(s) or case study without prior written approval.
 11. The above guidelines also apply to charts as they are graphical representations of cross-tabulations. In addition, graphical outputs (e.g., scatterplots, box plots, plots of residuals) should adhere to the above guidelines.
- G. That if the identity of any Private Person should be discovered, then:
1. No use will be made of this knowledge;
 2. <REPOSITORY ACRONYM> will be advised of the incident within five (5) business days of discovery of the incident;
 3. The information that would identify the Private Person will be safeguarded or destroyed as requested by <REPOSITORY ACRONYM>; and
 4. No one else will be informed of the discovered identity.
- H. Unless other provisions have been made with <REPOSITORY ACRONYM>, all originals and copies of the Restricted Data, on whatever media, shall be destroyed on or before completion of this Agreement or within 5 days of written request from <REPOSITORY ACRONYM>. Investigator will complete and notarize an Affidavit of Destruction, attesting to the destruction of the Restricted Data. Investigators requiring the Restricted Data beyond the completion of this Agreement should submit a request for continuation three months prior to the end date of the agreement. This obligation of destruction shall not apply to Investigator’s scholarly work based upon or that incorporates the Restricted Data.
- I. That any books, articles, conference papers, theses, dissertations, reports, or other publications that employed the Restricted Data or other resources provided by <REPOSITORY ACRONYM> reference the bibliographic citation provided by <REPOSITORY ACRONYM> and be reported to <REPOSITORY ACRONYM>.
- J. To provide annual reports to <REPOSITORY ACRONYM> staff (through <REPOSITORY ACRONYM>’s online data access request system), which include:

1. A copy of the annual IRB approval for the project described in the Research Description;
 2. A listing of public presentations at professional meetings using results based on the Restricted Data or Derivatives or analyses thereof;
 3. A listing of papers accepted for publication using the Restricted Data, or Derivatives or analyses thereof, with complete citations;
 4. A listing of Research Staff using the Restricted Data, or Derivatives or analyses thereof, for dissertations or theses, the titles of these papers, and the date of completion; and
 5. Update on any change in scope of the project as described in the Research Description.
- K. To notify <REPOSITORY ACRONYM> of a change in institutional affiliation of the Investigator, a change in institutional affiliation of any Research Staff, or the addition or removal of Research Staff on the research project. Notification must be in writing and must be received by <REPOSITORY ACRONYM> at least six (6) weeks prior to the last day of employment with Institution. Notification of the addition or removal of Research Staff on the research project shall be provided to <REPOSITORY ACRONYM> as soon as reasonably possible. Investigator's separation from Institution terminates this Agreement.
- L. Upon Investigator's change in institutional affiliation, all electronic and paper Restricted Data will be securely destroyed with a notarized affidavit of destruction submitted to <REPOSITORY ACRONYM>. <REPOSITORY ACRONYM> will, at the request and cost of Investigator, store these files and transfer them to Investigator's new Institution upon submission and approval of an <APPLICATION NAME> by the new Institution. Although the Restricted Data will be stored in a secure location, <REPOSITORY ACRONYM> assumes no responsibility for the Restricted Data or associated files and Institution and Investigator shall not be liable for any damages arising from any suits or claims arising from the storage of the Restricted Data or associated files by <REPOSITORY ACRONYM>. <REPOSITORY ACRONYM> makes no guarantees and provides no warranty that the exact same Restricted Data or associated files can be or will be provided to Investigator after such storage, or that any files or Restricted Data forwarded to Investigator after such storage will be free from defect or fit for any particular purpose.
- M. That use of the Restricted Data will be consistent with the Institution's policies regarding scientific integrity and human subject's research.
- N. To respond fully and in writing within ten (10) working days after receipt of any written inquiry from <REPOSITORY ACRONYM> regarding compliance with this Agreement.

VII. Violations of this Agreement

- A. The Institution will investigate allegations by <REPOSITORY ACRONYM> or other parties of violations of this Agreement in accordance with its policies and procedures on scientific integrity and misconduct. If the allegations are confirmed, the Institution will treat the violations as it would violations of the explicit terms of its policies on scientific integrity and misconduct.

- B. In the event of a breach of any provision of this Agreement, Institution shall be responsible to promptly cure the breach and mitigate any damages. The Institution hereby acknowledges that any breach of the confidentiality provisions herein may result in irreparable harm to <REPOSITORY ACRONYM> not adequately compensable by money damages. Institution hereby acknowledges the possibility of injunctive relief in the event of breach, in addition to money damages. In addition, <REPOSITORY ACRONYM> may:
1. Terminate this Agreement upon notice and require return of the Restricted Data and any derivatives thereof;
 2. Deny Investigator future access to Restricted Data; and/or
 3. Report the inappropriate use or disclosure to the appropriate federal and private agencies or foundations that fund scientific and public policy research.
 4. Such other remedies that may be available to <REPOSITORY ACRONYM> under law or equity, including injunctive relief.
- C. Institution agrees, to the extent not prohibited under applicable law, to indemnify the Regents of the University of Michigan from any or all claims, losses, causes of action, judgments, damages, and expenses arising from Investigator's, Research Staff's, and/or Institution's use of the Restricted Data, except to the extent and in proportion such liability or damages arose from the negligence of the Regents of the University of Michigan. Nothing herein shall be construed as a waiver of any immunities and protections available to Institution under applicable law.
- D. In the event of a violation, the Investigator must:
1. Notify <REPOSITORY ACRONYM> within five (5) business days;
 2. Stop work with the Restricted Data immediately;
 3. Submit a notarized affidavit acknowledging the violation to <REPOSITORY ACRONYM>;
 4. Inform the Representative of Institution of the violation and review security protocols and disclosure protections with them.
 - i. The Representative of Investigator's Institution must submit an acknowledgment of the violation and security protocols and disclosure protections review to <REPOSITORY ACRONYM>; and
 5. Reapply for access to the Restricted Data.

VIII. Confidentiality

To the extent the Restricted Data are subject to a Certificate of Confidentiality, the Institution is considered to be a contractor or cooperating agency of <REPOSITORY ACRONYM>; as such, the Institution, the Investigator, and Research Staff are authorized to protect the privacy of the individuals who are the subjects of the Restricted Data by withholding their identifying characteristics from all persons not connected with the conduct of the Investigator's research project. "Identifying characteristics" are considered to include those data defined as confidential under the terms of this Agreement.

IX. Incorporation by Reference

All parties agree that the information entered into the <ONLINE APPLICATION>, including the Data Security Plan, IRB approval, and any Supplemental Agreements and Confidentiality Pledges, are incorporated into this Agreement by reference.

X. Miscellaneous

- A. All notices, contractual correspondence, and return of Restricted Data under this Agreement on behalf of the Investigator shall be made in writing and delivered to the address below:

<REPOSITORY ACRONYM>
<REPOSITORY MAILING ADDRESS>
-or-
<REPOSITORY EMAIL ADDRESS>

- B. This agreement shall be effective for 24 months from execution or until the IRB expires.
- C. The respective rights and obligations of <REPOSITORY ACRONYM> and Investigator, Research Staff, and Institution pursuant to this Agreement shall survive termination of the Agreement.
- D. This Agreement and any of the information and materials entered into the <APPLICATION NAME> may be amended or modified only by the mutual written consent of the authorized representatives of <REPOSITORY ACRONYM> and Investigator and Institution. Both parties agree to amend this Agreement to the extent necessary to comply with the requirements of any applicable regulatory authority.
- E. The Representative of the Institution signing this Agreement has the right and authority to execute this Agreement, and no further approvals are necessary to create a binding agreement.
- F. The obligations of Investigator, Research Staff, and Institution set forth within this Agreement may not be assigned or otherwise transferred without the express written consent of <REPOSITORY ACRONYM>.

**Investigator and Institutional
Signatures**

Read and Acknowledged by:
Investigator

Institutional Representative

SIGNATURE

DATE

SIGNATURE

DATE

NAME TYPED OR PRINTED

NAME TYPED OR PRINTED

TITLE

TITLE

INSTITUTION

INSTITUTION

BUILDING ADDRESS

BUILDING ADDRESS

STREET ADDRESS

STREET ADDRESS

CITY, STATE ZIP

CITY, STATE ZIP