Tiered Access to Research Data for Secondary Analysis

John E Marcotte, Sarah Rush & Kelly Ogden-Schuette
University of Michigan

As the richness of research data about humans for secondary analysis has grown, disclosure risk has also increased. Research data have expanded in their gradation of the risks associated with both re-identification and potential harm[1], which has created a need for multiple levels of access controls beyond public and restricted access. Moreover, greater awareness of privacy, heightened standards for research ethics, and new laws necessitate additional gradations of access. Public-access data have typically been available for download from websites with minimal conditions on how data may be used while restricted-access data usually require an application and formal authorization process.

Throughout this paper, we characterize research data about humans as information for producing summary results such as contingency tables, means and medians as well as regression coefficients and transformations such as odds ratios and relative risks. These results must meet disclosure protection thresholds for cell sizes in tables, and sample sizes for regressions, as well as suppression of certain variables and disallowed sub-samples. Although research data may contain information about individuals and organizations, research data are not intended for identifying individuals or organizations. Whether data contain information about individuals or aggregates, the purpose of producing summary results is the same. According to the National Institutes of Health (NIH) "Common Rule", scientific research is:

> "A systematic investigation including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge." (Code of Federal Regulations CFR 46.102 Common Rule)

Research data are for conducting scientific inquiry and are for the calculation of summary measures only and must not be accessed to identify a specific individual, organization, or community.

Data stewards and researchers have a legal and ethical obligation to protect the identities of participants in research data. Methods of access to research data must ameliorate threats while at the same time avoiding excessive barriers to scientific inquiry. Security controls must align with the re-identification and potential harm risk in the data. Although bona fide researchers very rarely intentionally allow data to leak, inadvertent breaches still occur. Researchers must follow specified protocols and when necessary, the rules must have checks for compliance. The paradigm of classifying data as public-access or restricted-access is no longer sufficient. Access to research data requires more nuance to ensure the protection of human subjects.

Background

Over the last 25 years, several articles and reports have developed frameworks for providing access to research data. In a 2002 report entitled *Restricted Access Procedures*[2], the Confidentiality and Data Access Committee identify Research Data Centers (RDC) as a primary

[1] AHRQ Common Formats https://pso.ahrq.gov/common-formats/overview

[2] Federal Committee on Statistical Methodology: Confidentiality and Data Access Committee. (2002, April). *Restricted Access Procedures*. https://nces.ed.gov/FCSM/pdf/CDAC_RAP.pdf

method for providing access to restricted-access research data. The report also discusses remote access and online query systems as alternatives to RDC. At the time of this report, remote access systems were still in their infancy. Several years later, Kinney, Karr and Gonzales (2010)[3] discuss direct access through RDC and licensed access for researchers to analyze data on their own computers. Kinney, et al. also propose using tabular and synthetic data to mitigate disclosure risk. More recently, Desai, Ritchie, and Welpton (2016)[4] describe the Five Safes framework for data access. The framework, based on aspects related to the project, people, data, settings, and output, can be a basis for designing tiered access. Desai, et al. describe a data access spectrum. Our approach builds on the concept of safe locations by specifying access tiers in terms of controls to ensure compliance with protocols.

More than 1,800 public research data repositories are currently available to academia, government, and business.[5] Several of these repositories have established different levels of access that have some overlap with the seven tiers we propose in this paper. These levels usually focus on technology. While some repositories have offered tiered access, our unique contribution is how we define the tiers in terms of both human and technical controls to prevent the release of disclosive information.

Dataverse[6] offers tiered access but only specifies additional technical controls such as encryption and two-factor authentication. To our knowledge, Dataverse does not have a system in place to review output.

Guidance for Controllers on Data Security (February 2020)[7] from the **Irish Social Science Data Archive (ISSDA)[8]** is an example of how repositories approach both physical security and the human factor. The Physical Security requirements overlap with many other repositories while the discussion of the human factor does not suggest controls beyond training, accountability, and continuity.

---

[3] Kinney, S. K., Karr, A. F., & Gonzalez Jr., J. F. (2010). Data Confidentiality: The Next Five Years Summary and Guide to Papers. *Journal of Privacy and Confidentiality*, *1*(2). https://doi.org/10.29012/jpc.v1i2.569

[4] Desai, T., Ritchie, F., & Welpton, R. (n.d.). Five Safes: designing data access for research. In *Economics Working Paper Series 1601*. University of the West of England, Bristol. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf

[5] Crosas, M. (n.d.). *CIO Review: Cloud Dataverse: A Data Repository Platform for the Cloud*. https://openstack.cioreview.com/cxoinsight/cloud-dataverse-a-data-repository-platform-for-the-cloud-nid-24199-cid-120.html

[6] *The Dataverse Project*. (n.d.). https://dataverse.org/

[7] Data Protection Commission (Ireland). (2020, February). *Guidance for Controllers on Data Security*. https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Data_Security_Guidance_Feb20.pdf

[8] *Irish Social Science Data Archive (ISSDA)*. (n.d.). University College Dublin. https://www.ucd.ie/issda/

Slavkovic, Kinney, and Karrin writing in *Chance* (2013)[9] cite the **National Opinion Research Center (NORC)[10]** Data Enclave as an example of an online data enclave. They describe both technical and human controls including how researchers access data over an encrypted connection and are unable to transfer any data, even via copy and paste, to their local computer. Moreover, output must be reviewed before release to researchers. They also discuss the cost of operations. Slavkovic, et al. also mention Census Research Data Centers, but they do not indicate what additional security controls the Census Research Data Centers have over an online enclave.

According to Thissen and Mason in *Health Systems* (2019)[11], security controls for research data depend both on the sensitivity of the information and on regulations, requirements, or ethical constraints. Compliance with regulations is a key aspect of specifying controls. Thiessen and Mason do not discuss how compliance is ensured.

Horton, Perry, and Bishop (2020)[12] present three tiers: (1) Open, (2) Accountable and (3) Controlled. The term restricted applies to both accountable and controlled. In our view, three levels are not sufficient for providing access to research data because the gradation of risk requires more options.

In "Sharing Confidential Data for Research Purposes A Primer*,"* Reitera and Kinney (2011)[13] identify two primary restricted-access methods employed by most data stewards, including government agencies and individual investigators: licensing agreements and restricted-data centers. These access methods correspond to some of the tiers that we discuss below. Reitera and Kinney do acknowledge online enclaves such as NORC's for providing access, but do not specifically discuss other tiers.

Any discipline that analyzes research data must be concerned with security. Social scientists as well as medical and public health researchers are all dealing with how to provide appropriate access to research data. Lawyers and computing professionals tend to approach access to research data as a problem of licenses and waivers. (Institutional) Data Use Agreements (DUA) between organizations, (Individual) Terms of Use (TOU) accepted by researchers are common nomenclature. While these descriptors may have overlapping definitions, repositories often apply them in different circumstances. A valid agreement or license is required to access the research data. Waivers typically refer to unrestricted access. Computing professionals often focus

---

[9] Slavkovic, A., Kinney, S., & Karr, A. (2013, August 2). O Privacy, Where Art Thou? *Chance*, *24*(4), 41-45. https://doi.org/10.1080/09332480.2011.10739886

[10] *National Opinion Research Center*. (n.d.). NORC at the University of Chicago. https://www.norc.org

[11] Thissen, M. R., & Mason, K. M. (2019, April 15). Planning security architecture for health survey data storage and access. *Health Systems*, *9*(1), 57-63. https://doi.org/10.1080/20476965.2019.1599702

[12] *Open where possible, closed if necessary: reforming access categories for social science data archives* [Presentation]. (2020, February 17). International Digital Curation Conference 2020 (IDCC20), Dublin, Ireland. https://doi.org/10.5281/zenodo.3670943

[13] Reiter, J. P., & Kinney, S. K. (2011, September). Commentary: Sharing Confidential Data for Research Purposes: A Primer. *Epidemiology*, *22*(5), 632-635. https://doi.org/10.1097/EDE.0b013e318225c44b

on physical security, authentication, authorization, audit, and encryption. Training is frequently the specified human control. In our paradigm, we specify 10 security controls to ensure that disclosive information is not released.

Some paradigms treat "trustworthiness" as a continuum instead of as a minimum requirement. In our proposed approach, researchers must meet minimum requirements to access restricted data. A higher trust score does not entitle the researcher to relax security protocols nor automatically access other restricted-use data.

Seven Tiers of Access

In this paper, we propose seven tiers of access to research data. Each tier adds requirements that are necessary to mitigate disclosure risk and affirm appropriate management of the data. Improper handling of the data include attempts to find a specific individual or household or failure to follow disclosure protection rules for data and output included in papers and presentations. By establishing a ladder of access conditions, each higher tier meets and exceeds the requirements of the lower tiers. While the highest tier meets all requirements, this tier will impede legitimate research for most data. The challenge for repositories is to provide access in a manner that promotes research while specifying security that provides appropriate protections against the risks of re-identification and harm. The tiers operationalize risk management options. The requirements of the research data determine the appropriate tier. Researchers must qualify for access, and all access is through that tier or a more restricted tier only. Although data repositories have provided some of these tiers, all seven tiers are necessary to meet the growing gradation of risk in research data.

The tiers of access range from 0-Unrestricted to 6-Batch. At all tiers, research data, by definition, are not for identifying individuals or organizations. While researchers promise to follow protocols at all tiers, each tier adds a control that ensures compliance. As the risk in data increases, pledges are insufficient to protect human subjects. While intentional non-compliance is relatively rare, researchers primarily focus on their scientific inquiry and may inadvertently fail to follow rules created to safeguard the data.

The seven tiers are:

- 0-Unrestricted (public use)
- 1-Registered
- 2-Approved
- 3-Local
- 4-Remote
- 5-Vault
- 6-Batch (all controls)

While all tiers of access typically require that researchers agree to protect human subjects and only publish non-disclosive results, data accessed through tier "0-Unrestricted" may usually be downloaded from a website after researchers agree to only analyze the data for research and not to re-identify specific people in the data. Data available as "0-Unrestricted" do not enable re-identification; nevertheless, by accepting the terms of use, researchers agree to not even try. The descriptions of tiers 1 and 2 are out of order because they are neither pubic nor restricted.

Tiers 1 and 2 are for situations where researchers still need to apply for access, but do not need an institutional DUA.

Tiers 3 through 6 are typically grouped together under the classification of restricted; these tiers require an application. Applicants must obtain approval for the research from an Institutional Review Board (IRB) or comparable ethics panel and submit a data security plan to protect the data from leakage. Applicants may also be required to obtain training. To access data in these tiers, the universities or organizations of these applicants must enter into a Data Use Agreement (DUA). This institutional agreement specifies that the applicant's research may be conducted under the auspices of the university or organization. Moreover, the DUA requires the university or organization to take appropriate action including research misconduct proceedings if a protocol is violated.

In sum, for tiers 0,1 and 2, researchers can agree to Terms of Use (TOU) while for tiers 3,4,5,6, researchers, and their organizations must agree to terms.

The seven tiers build on 10 controls for accessing research data. The specification of these controls for delineating the access tiers is unique to this paper. These 10 controls protect against the disclosure, re-identification, and harm risks associated with a particular dataset. The regulations and security controls form a ladder and allow higher tiers to build on the protocols of lower tiers:

- Application: (a) must apply for access or (b) no application is necessary.
- Approval: (a) must receive approval for access or (b) no approval is necessary.
- Agreement: (a) institution (university or organization) and individual researcher or (b) individual researcher only.
- Period of Access: (a) a specified period only or (b) unlimited time.
- Research Location: (a) specified approved locations only or (b) any location.
- Encryption: (a) encrypted at rest and in transit or (b) clear text is sufficient.
- Internet: (a) both outbound and inbound internet connection are blocked or (b) access to the Internet is allowed.
- Output: (a) output must be reviewed for adherence to disclosure protection rules either by an authorized reviewer or by the researcher or (b) output does not require review.
- Proctor: (a) an authorized guard or monitor must be present during data access or (b) no proctor is required.
- View Data: (a) researchers can only view approved summary results from data not the actual research data or (b) researchers may view the research data

The following chart shows how the access tiers mesh with the controls. At each level, researchers must agree and comply with all regulations. As risks increase, researcher agreement is not sufficient and technical configurations must prevent researchers from inadvertent data disclosure. While researchers agree to follow all conditions, each tier adds a layer of security that ensures researcher compliance. These extra security layers are an impediment to research and should only be implemented when risks of re-identification and harm necessitate.

Access Tier by Control

	Tier	Description	Application	Approval	Agreement	Period of Access	Research Location	Encryption	Internet	Output	Proctor	View Data
Public	0-Unrestricted	researcher may download	none	none	Researcher	No limit	public or private	not required	allowed	not vetted	not monitored	allowed
	1-Registered	researcher must provide additional info such as research purpose before download	submit information	none	Researcher	No limit	public or private	not required	allowed	not vetted	not monitored	allowed
	2-Approved	researcher must be approved before download	must apply	approved	Researcher & Advisor	Limited	private	at rest in transit	allowed	not vetted	not monitored	allowed
Restricted	3- Local	researcher receives data with approved security plan	must apply	approved	Researcher & Institution	Specified period	private	at rest, real-time in transit	blocked	self-vetted	not monitored	allowed
	4-Remote	researcher comes to data electronically with approved security plan	must apply	approved	Researcher & Institution	Specified period	private	at rest in transit	blocked except session	externally vetted	not monitored	allowed
	5-Vault	researcher comes to data in person with pre-approved materials	must apply	approved	Researcher & Institution	Specified Period	private	at rest in transit	blocked	externally vetted	watched during access	allowed
	6-Batch	researchers cannot access the data researchers can only access summary results	must apply	approved	Researcher & Institution	Specified period	private	at rest	only batch submissions	externally vetted	monitored batch jobs	not allowed

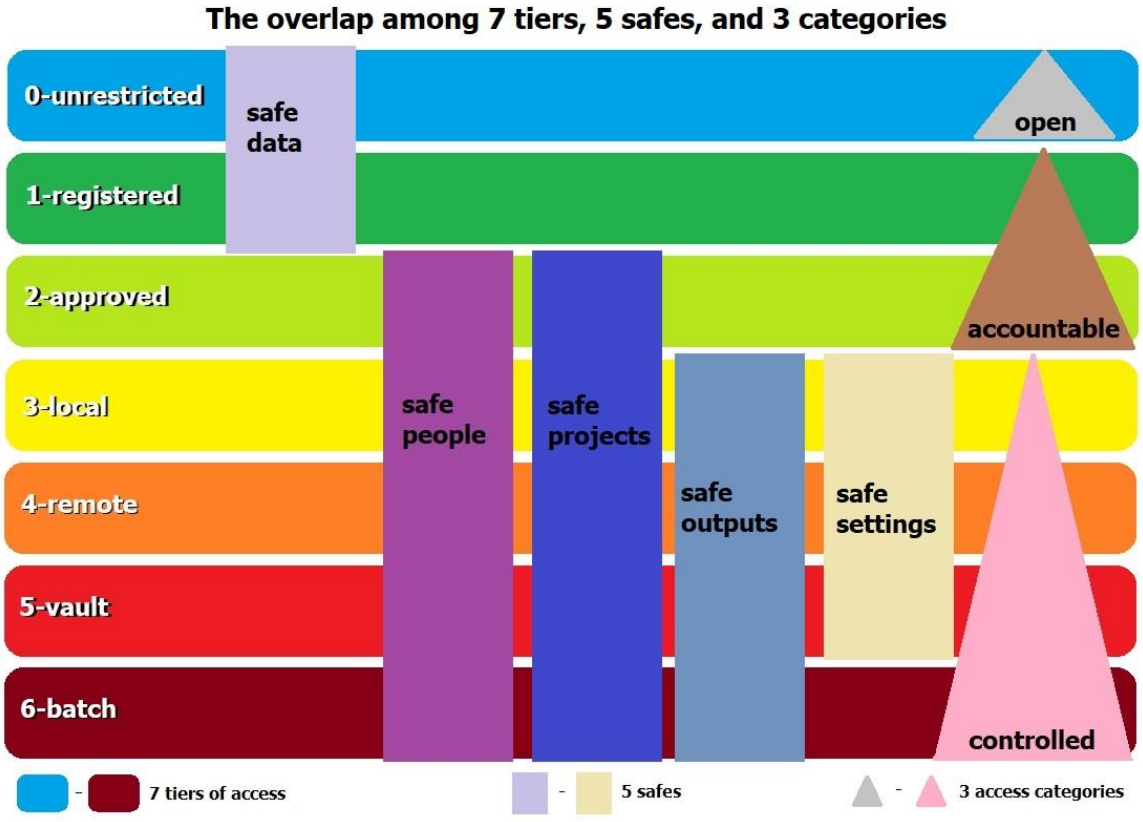
More details about each security control follow:

- **Application** to access the data. Only data in the 0-Unrestricted tier do not require an application. For data in tier 3-Local and above, some repositories require researchers to submit IRB or ethics panel approval, a security plan, and confidentiality pledges. Furthermore, for data in tiers 3 and above, only researchers who can serve as Principal Investigators (PI) may apply. Those researchers who are not PI-eligible, such as graduate students, may analyze the data only under the supervision of a PI.
- **Approval** to access the data. While tier 1-Registered-use requires researchers to submit information about research plans, only tier 2-Approved and above require submitted information to be reviewed and approved before access is granted. For tier 1, access to the data is provided immediately after the required information is submitted.
- **Agreement** is whether the researcher only or researcher and their institutional (or university or organizational) representative must sign the agreement to access the data. At tier 2-Approved and below, researchers can obtain access through their own agreement. For tier 3-Local and above, a university or institutional representative with authority to obligate the researcher's organization as well as the researcher must agree to and sign the data use agreement.
- **Period of Access** is either unlimited or is limited to a specified period. Tier 3-Local and above are only accessible until an end date. Tier 2-Approved may have limits on how long researchers can access the data. Tiers 0 and 1 allow for unlimited access. Agreements that require a university of institutional signature are always time bound.
- **Research Location** is where the data will be viewed. The research office has the client computer which may store the data or be a portal to a server where the data are stored. For tier 2-Approved and above, the client location must be private and specified accessing the data from a library or café is not permitted. A private location prevents inadvertent eavesdropping of the computer screen. A private home office is permitted as long as the location is approved.
- **Encryption** alleviates the ramifications of theft, loss of data, or interception. . Research data that require approval (Tiers 2 and above) must implement encryption in transit and at rest.
- **Internet** concerns both inbound and outbound network traffic on the machine being used to access the data. For tier 3-Local and above, access to the internet must be blocked. For tier 3, data must be analyzed and stored on a computer without an internet connection. The requirement is usually met by the researcher agreeing to analyze the data on a standalone non-networked computer. For tier 4-Remote, while the server allows inbound session connections only; outbound connections and other types of inbound connections such as SSH and HTTP, are not allowed. For tier 4-Remote, the systems administrator configures the system to block internet access. The purpose of blocking the internet is to prevent researchers from inadvertently copying files to unauthorized locations (typically through drag and drop). An acceptable configuration implements a two-step process of copying files off the computer with the research data. Blocking the internet also prevents the computer from being compromised and having any file stolen.
- **Output** must be vetted for compliance with disclosure protection rules such as minimum cell counts and minimum subsample sizes for regressions. Tier 3-Local and above require vetting, but all levels require compliance with rules about output. While tier 3 authorizes self-vetting, tier 4-Remote and above require trained personnel who are not part of the research project to review files before release in addition to the project team. Researchers may view and analyze data; however, they must submit output for review before export from the computer system where the data are accessed. In tiers 0-2, the research data should not contain sufficient information to produce disclosive outputs.

- **Proctor** is a guard who monitors researchers while accessing the data. For all tiers, researchers are not allowed to look up specific respondents in the data nor transcribe data points. For Tier 5-Vaultand above, researchers may only access the data in the presence of a proctor. Tier 5 and above prevent unsanctioned use of the data by taking unauthorized notes or files.
- **View Data** controls whether researchers can see the micro data or only summary results. At Tier 6-Batch, researchers cannot view the micro data. While at all tiers, researchers agree to not attempt to re-identify or look up a particular respondent, at Tier 6, researchers are prevented from even accessing the micro data, so any re-identification or lookups are impossible.

Seven Tiers, Five Saves, Three Access Catories

The sever tiers overlap with the five safes of Ritchie, et al. and the three access categories of Horton, et al. Horton's three categories are a combination of multiple tiers and form a similar hierarchy. Ritchies's five safes overlap multiple tiers and is a different conceptualization. Safe people and safe projects are aspects of approved people and projects. Safe outputs correspond to tiers where output must be reviewed for compliance with disclosure protection rules. The following chart illustrates the overlap.



Let us consider the implementation of the seven proposed tiers. All research data regardless of tier are for the calculation of summary measures only and must not be used to locate an individual, organization, or community. Higher levels build on lower tiers by adding more security controls.

0-Unrestricted

Public-access research data are typically available for download from websites or via an API (Application Programming Interface). These data are available without restrictions on access. Disclosure and harm risks are negligible; nevertheless, the data are for research only and must not be analyzed to locate an individual, organization, or community. Unrestricted research data are also labeled public-use and open-data. In many situations, public-use research data may be downloaded anonymously. The researcher who downloads the data can agree to the terms of use without an institutional signature.

- *Example:* Baby's First Years (BFY), New York City, New Orleans, Omaha, and the Twin Cities 2018-2021[14] has data of this type.
- *Implementation:* Website and bandwidth to handle download demand.
- *Weakness:* Data might still have hidden risks.
- *Impediment to research:* Data may not contain sufficient information for some types of analysis.

1-Registered

Registered research data are also typically available for download from websites. Disclosure and harm risks are very low. Unlike public-access data, registered data may not be downloaded anonymously. To register, researchers must provide valid contact information and a research purpose; however, download of the data does not require approval. The researcher who downloads the data can agree to the terms of use without an institutional signature

- *Example study:* National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [Public Use][15] requires registration from anyone downloading the ll data.
- *Implementation:* Registration system to collect information. Website and bandwidth to handle download demand.
- *Weakness:* Researchers could provide inaccurate information.
- *Impediment to research:* Researcher must provide information to access data.

2-Approved

Approved research data require registration and approval before download. The persons approving access will vary. The data repository may approve applications or data collectors may want to perform approvals. While these data have low disclosure risk, they may contain information that could be construed as being sensitive. Because the data are only available upon approval, researchers may have to implement additional safeguards for these data such as encryption. The researcher may only be allowed to access the data in a private setting. The researcher who downloads the data can agree to the terms of use. In some cases, a department chair or graduate advisor may need to supervise the research. The data collector in consultation with the data repository determines requirements.

---

[14] Magnuson, Katherine A., Noble, Kimberly, Duncan, Greg J., Fox, Nathan A., Gennetian, Lisa A., Yoshikawa, Hirokazu, and Halpern-Meekin, Sarah. Baby's First Years (BFY), New York City, New Orleans, Omaha, and Twin Cities, 2018-2019. Inter-university Consortium for Political and Social Research [distributor], 2020-11-16. https://doi.org/10.3886/ICPSR37871.v2

[15] Harris, Kathleen Mullan, and Udry, J. Richard. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2022-02-09. https://doi.org/10.3886/ICPSR21600.v24

- *Example:* Some Panel Study of Income Dynamics (PSID)[16] and Health and Retirement Study (HRS)[17] data fall into this category.
- *Implementation:* Application system with encrypted download.
- *Weakness:* Researchers could leak data inadvertently.
- *Impediment to research:* Researchers must apply for access to research data and wait for approval.

3-Local

Research data in this tier are restricted; however, access to the data is at the researcher's local university or organization. These data have a higher risk of re-identification and harm if disclosure occurs. Local data require an application and approval, but unlike 2-Approved, an institutional representative must sign the DUA in addition to the researcher. In the DUA, the institutional or organizational representative must verify that the researcher is qualified and affiliated with the institution. Moreover, the institution must have rules governing research misconduct and must agree to invoke these protocols if an infraction occurs. Qualified researchers must be PI eligible to access these data; other researchers and students must work under the supervision of a qualified researcher. Analysis of these data require IRB approval and confidentiality pledges from personnel who can access the data. In addition to whole disk encryption, the data must reside on a standalone (non-networked) computer in a private office. The researcher must also agree to abide by disclosure protection rules and must self-review articles and output for compliance.

- *Example study:* NICHD Study of Early Child Care and Youth Development (SECCYD)[18] has data in this tier.
- *Implementation:* Standalone (non-networked computer) in a locked private office. Some organizations may have an acceptable server set up.
- *Weakness:* The research data with re-identification and harm risks are not under the control of the repository. Unauthorized access is possible.
- *Impediment to research:* Difficult to collaborate with a research team. Universities and organizations may be reluctant to permit a non-networked computer. Researchers may not have extra funds to buy a another computer that is dedicated to a single project.

4-Remote

Research data in tier 4-Remote have the same application requirements as 3-Local. Instead of researchers analyzing the data on systems at the local organization, they access the data through encrypted connections to a "Virtual Data Enclave" or "Virtual Research Data Center." These data

---

[16] Johnson, David S., Freedman, Vicki A., Sastry, Narayan, McGonagle, Katherine A., Brown, Charles, Fomby, Paula, … Stafford, Frank P. Panel Study of Income Dynamics (PSID): Main Interview, 1968-2015. Inter-university Consortium for Political and Social Research [distributor], 2018-10-04. https://doi.org/10.3886/ICPSR37142.v1

[17] *Health and Retirement Study, public use dataset*. (n.d.). Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). https://hrsdata.isr.umich.edu/data-products/public-survey-data

[18] United States Department of Health and Human Services. National Institutes of Health. Eunice Kennedy Shriver National Institute of Child Health and Human Development. (n.d.). *NICHD Study of Early Child Care and Youth Development: Phases I-IV [United States]*. Inter-university Consortium for Political and Social Research [distributor]. https://www.icpsr.umich.edu/web/DSDR/series/233

may have higher re-identification and harm risk. In some cases, these data may be linked with other information such as geographic contextual variables. Data at this level are stored in an enclave and cannot be downloaded to the local computer, so that the repository retains control over access to the data. Researchers must review their output for compliance with disclosure protection rules; however, trained repository staff must also vet the output. Only files that meet disclosure protection requirements are released out of the enclave. Restrictions on additional data that can be linked can also be enforced. Besides allowing a secondary level of output vetting, enclaves offer the additional benefit of enabling research teams to collaborate on the analysis of data with disclosure risk. Enclaves are fast becoming the preferred method for restricted data access and eventually will subsume Tier 3-Local.

- *Example:* The restricted Los Angeles Family and Neighborhood Survey (L.A.FANS)[19] data are in this tier.
- *Implementation:* Terminal Server or Virtual Desktop Infrastructure that prevents files from being copied off the server or VDI. ICPSR[20], NORC[21], and Survey Research Center[22] at the Institute for Social Research[23] have enclaves in production.
- *Weakness:* Researchers could still transcribe information from the screen.
- *Impediment to research:* Researchers must wait for the release of results. Available software may be limited. The computation power of the virtual machines may not be sufficient for some research.

5-Vault

Vault protocols add a proctor or observer to the security requirements. A "vault" is a locked room where data can only be accessed in this locked room in the presence of an observer. The proctor checks that only approved information is extracted from the data and taken out of the 'vault." As with tiers 3-Local and 4-Remote, this level requires an application and approval. These data typically have even higher re-identification and harm risks.

- *Example:* Videos are an example of data in this tier. Most videos are high disclosure risk. As with data in 4-Remote, all output and files are reviewed before release.
- *Implementation:* Locked room with proctor. Federal Statistical Research Data Centers[24] have implemented this level of security. Some research centers have had vault rooms for accessing restricted data.
- *Weakness:* Researchers could still look up an individual record.
- *Impediment to research:* Accessing the data requires travel to the vault location and an appointment. Repository staff must also allocate time to work in the vault to serve as proctors.

---

[19] Pebley, A. R., & Sastry, N. (n.d.). *Los Angeles Family and Neighborhood Survey (L.A.FANS), Waves 1-2: Restricted Data Versions 1-3; Restricted Neighborhood Observations Data*. Inter-university Consortium for Political and Social Research [distributor]. https://www.icpsr.umich.edu/web/DSDR/series/846

[20] *Inter-university Consortium for Political and Social Research*. (n.d.). ICPSR. http://icpsr.umich.edu

[21] *National Opinion Research Center*. (n.d.). NORC at the University of Chicago. https://www.norc.org

[22] *Survey Research Center*. (n.d.). https://www.src.isr.umich.edu/

[23] *Institute for Social Research*. (n.d.). https://isr.umich.edu/

[24] *Federal Statistical Research Data Centers*. (n.d.). Census Bureau. https://www.census.gov/about/adrm/fsrdc.html

6-Batch

This tier is for research data with the highest risks and provides the maximum protection since researchers are unable to view the micro data. Researchers are only allowed to see approved summary results.. Data in 6-Batch have both high sensitivity and high re-identification risks. Researchers must submit requests for regressions or crosstabs to the data repository. An automated system or repository staff generate the requested summary statistics. Only after the results are vetted for disclosure risk are the results released to researcher. Accessing these data requires an application and approval as well as institutional agreement. While this tier does not allow researchers to touch the micro data, this level has one advantage over 5-Vault in that it does not require travel.

- *Example: LISSY* at the Cross-national Data Center in Luxembourg[25] is an implementation of this tier. The retired *ANDRE* system at the National Center for Health Statistics (NCHS)[26] was also an example.
- *Implementation:* Batch system. A server with synthetic data and the software available in the batch system for testing programs will enable the system to run smoothly.
- *Impediment to research:* Without access to the data, analysis is cumbersome and requires much more time. Even though 6-Batch is more restrictive than 5-Vault, the tier does not require travel to a specific location.

Although each of these tiers may be available for different studies, our specification shows how each tier adds a control to ensure compliance. For research data with human subjects, tiered access is essential since data vary in their risks of re-identification and harm.

In conclusion, while tiered access to research data is not a new idea, more than two or three levels are needed to meet the diverse needs of the research community. In this paper, we propose seven tiers along with detailed descriptions of each tier as well as examples of data that fall within each tier. With these seven tiers of access, repositories can meet the needs of researchers while still providing appropriate protections for research data. The tiered approach enables repositories to require sufficient security controls without creating unnecessary impediments to research.

Bibliography

Crosas, M. (n.d.). *CIO Review: Cloud Dataverse: A Data Repository Platform for the Cloud*. https://openstack.cioreview.com/cxoinsight/cloud-dataverse-a-data-repository-platform-for-the-cloud-nid-24199-cid-120.html

Data Protection Commission (Ireland). (2020, February). *Guidance for Controllers on Data Security*. https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Data_Security_Guidance_Feb20.pdf

Desai, T., Ritchie, F., & Welpton, R. (n.d.). Five Safes: designing data access for research. In *Economics Working Paper Series 1601*. University of the West of England, Bristol. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf

[25] *LIS Cross-National Data Center in Luxembourg*. (n.d.). https://www.lisdatacenter.org/data-access/lissy/

[26] *NCHS - National Center for Health Statistics*. (n.d.). CDC. https://www.cdc.gov/nchs/index.htm

Federal Committee on Statistical Methodology: Confidentiality and Data Access Committee. (2002, April). *Restricted Access Procedures*. https://nces.ed.gov/FCSM/pdf/CDAC_RAP.pdf

Federal Statistical Research Data Centers. (n.d.). Census Bureau. https://www.census.gov/about/adrm/fsrdc.html

Harris, Kathleen Mullan, and Udry, J. Richard. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor], 2022-02-09. https://doi.org/10.3886/ICPSR21600.v24

Health and Retirement Study, public use dataset. (n.d.). Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). https://hrsdata.isr.umich.edu/data-products/public-survey-data

Institute for Social Research. (n.d.). https://isr.umich.edu/

Inter-university Consortium for Political and Social Research. (n.d.). ICPSR. http://icpsr.umich.edu

Irish Social Science Data Archive (ISSDA). (n.d.). University College Dublin. https://www.ucd.ie/issda/

Johnson, David S., Freedman, Vicki A., Sastry, Narayan, McGonagle, Katherine A., Brown, Charles, Fomby, Paula, … Stafford, Frank P. Panel Study of Income Dynamics (PSID): Main Interview, 1968-2015. Inter-university Consortium for Political and Social Research [distributor], 2018-10-04. https://doi.org/10.3886/ICPSR37142.v1

Kinney, S. K., Karr, A. F., & Gonzalez Jr., J. F. (2010). Data Confidentiality: The Next Five Years Summary and Guide to Papers. *Journal of Privacy and Confidentiality*, *1*(2). https://doi.org/10.29012/jpc.v1i2.569

LIS Cross-National Data Center in Luxembourg. (n.d.). https://www.lisdatacenter.org/data-access/lissy/

Magnuson, Katherine A., Noble, Kimberly, Duncan, Greg J., Fox, Nathan A., Gennetian, Lisa A., Yoshikawa, Hirokazu, and Halpern-Meekin, Sarah. Baby's First Years (BFY), New York City, New Orleans, Omaha, and Twin Cities, 2018-2019. Inter-university Consortium for Political and Social Research [distributor], 2020-11-16. https://doi.org/10.3886/ICPSR37871.v2

National Opinion Research Center. (n.d.). NORC at the University of Chicago. https://www.norc.org

NCHS - National Center for Health Statistics. (n.d.). CDC. https://www.cdc.gov/nchs/index.htm

Open where possible, closed if necessary: reforming access categories for social science data archives [Presentation]. (2020, February 17). International Digital Curation Conference 2020 (IDCC20), Dublin, Ireland. https://doi.org/10.5281/zenodo.3670943

Pebley, A. R., & Sastry, N. (n.d.). *Los Angeles Family and Neighborhood Survey (L.A.FANS), Waves 1-2: Restricted Data Versions 1-3; Restricted Neighborhood Observations Data*. Inter-university Consortium for Political and Social Research [distributor]. https://www.icpsr.umich.edu/web/DSDR/series/846

Reiter, J. P., & Kinney, S. K. (2011, September). Commentary: Sharing Confidential Data for Research Purposes: A Primer. *Epidemiology*, *22*(5), 632-635. https://doi.org/10.1097/EDE.0b013e318225c44b

Slavkovic, A., Kinney, S., & Karr, A. (2013, August 2). O Privacy, Where Art Thou? *Chance*, *24*(4), 41-45. https://doi.org/10.1080/09332480.2011.10739886

Survey Research Center. (n.d.). https://www.src.isr.umich.edu/

The Dataverse Project. (n.d.). https://dataverse.org/

Thissen, M. R., & Mason, K. M. (2019, April 15). Planning security architecture for health survey data storage and access. *Health Systems*, *9*(1), 57-63. https://doi.org/10.1080/20476965.2019.1599702

United States Department of Health and Human Services. National Institutes of Health. Eunice Kennedy Shriver National Institute of Child Health and Human Development. (n.d.). *NICHD Study of Early Child Care and Youth Development: Phases I-IV [United States]*. Inter-university Consortium for Political and Social Research [distributor]. https://www.icpsr.umich.edu/web/DSDR/series/233