Journal of Privacy and Confidentiality
Vol. 13 (2) 2023

NAHDAP-ICPSR

Submitted Nov 2022
Published Nov 2023

TIERED ACCESS TO RESEARCH DATA FOR SECONDARY ANALYSIS

JOHN E. MARCOTTE, SARAH RUSH, AND KELLY OGDEN-SCHUETTE

University of Michigan
e-mail address: jemarcot@umich.edu

University of Michigan

University of Michigan

Abstract. Research data have expanded in their gradation of the risks associated with both re-identification and harm, which has created a need for multiple levels of access controls beyond public and restricted access. Public-access data have typically been available for download from websites while restricted-access data usually require an application and formal authorization process. The old 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. In this paper, we describe seven tiers of access to research data. Each tier adds requirements that are necessary to mitigate disclosure risk and confirm appropriate management of the data. Improper handling of the data includes attempting to find a specific individual or household or failing 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.

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 control beyond only 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

Key words and phrases: Confidential data; Data reuse; Sensitive data.

[*] Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R24HD048404. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[1] See, for example, AHRQ Common Formats; https://pso.ahrq.gov/common-formats/overview.

JPrivacy Confidentiality

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.853

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, they are not intended for identifying those 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.[2]

Research data are for conducting scientific inquiry, 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 avoid excessive barriers to scientific inquiry. Security controls must align with the re-identification and potential harm risk in the data. Although 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 either public-access or restricted-access is no longer sufficient. Access to research data requires more nuance to ensure the protection of human subjects.

1. Background

Over the last 25 years, several articles and reports have developed frameworks for providing access to research data. In a 2002 report [4], the Confidentiality and Data Access Committee of the Federal Committee on Statistical Methodology (FCSM) identifies Research Data Centers (RDCs) as a primary method for enabling use of restricted-access research data. The report also discusses remote access and online query systems as alternatives to RDCs. At the time of this report, remote access systems were still in their infancy. Several years later, Kinney, Karr, and Gonzales [6] discuss direct access through RDCs 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 [3] 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. outline a data access spectrum. Our approach builds on the concept of safe locations by specifying access tiers in terms of controls meant to ensure compliance with protocols.

More than 1,800 public research data repositories are currently available to researchers in academia, government, and business [1]. Several of these repositories have established differing levels of access that have some overlap with the seven tiers we propose in this paper. Those 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.

[2] CFR 46.102.

For example, *Dataverse*[3] 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.

The Irish *Guidance for Controllers on Data Security* [2] is an example of how repositories approach both physical security and the human factor. The physical security requirements overlap those of many other repositories, while the discussion of the human factor does not suggest controls beyond training, accountability, and continuity.

Slavkovic, Kinney, and Karr [10] cite the National Opinion Research Center (NORC) 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 computers. Moreover, output must be reviewed before release to researchers. They also discuss the cost of operations. Slavkovic et al. mention Census Bureau RDCs, but they do not indicate what additional security controls the these RDCs have over an online enclave.

According to Thissen and Mason [11], security controls for research data depend not only on the sensitivity of the information, but also on regulations, requirements, and 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 [5] present three categories of access: (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.

Reiter and Kinney [9] identify two primary restricted-access methods employed by many 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. Reiter and Kinney acknowledge online enclaves such as that of NORC, 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 all deal 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 (DUAs) between organizations and (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 on physical security, authentication, authorization, audit, and encryption. Training is frequently the specified human control. In our paradigm, we specify ten 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 approach, researchers must meet minimum requirements to access restricted data. A higher trust score does not entitle the researcher to relaxed security protocols nor automatically to access other restricted-use data.

[3] https://dataverse.org/.

2. 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 includes attempts to find a specific individual,[4] as well as 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 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 promoting research while specifying security that provides appropriate protection against the risks of re-identification and harm. The tiers operationalize risk management options. The characteristics of the research data determine the appropriate tier. Researchers must qualify for access, and all access is only through that tier or a more restricted tier. 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, forbid identifying individuals. While researchers promise to follow protocols at all tiers, each tier adds a control that ensures compliance. As the risk in data increases, pledges alone are insufficient to protect human subjects. While intentional non-compliance is relatively rare, researchers focused primarily on their scientific inquiry may inadvertently fail to follow rules created to safeguard the data. The seven tiers are:

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

While all tiers of access require that researchers agree to protect data 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 individuals in the data. Data available as 0-Unrestricted do not enable re-identification; nevertheless, by accepting the TOU, researchers agree not even to try.

The descriptions of tiers 1 and 2 are out of order because they are neither public nor restricted. Tiers 1 and 2 are for situations in which 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. Applicants may also be required to obtain training. To access data in these tiers, employers (universities or other organizations) of these applicants must enter into a DUA with the owner or controller of the data. This institutional agreement specifies that the applicant's research will be conducted under the auspices of the organization.

[4] Although much of this paper applies to data on households or establishments, for simplicity we retain the language of individuals.

Moreover, the DUA requires the organization to take appropriate action if a protocol is violated, including research misconduct proceedings.

In sum, for tiers 0, 1 and 2, researchers can agree to TOU, while for tiers 3, 4, 5 and 6, researchers and their organizations must agree to terms set forth in a DUA.

The seven tiers build on ten controls for accessing research data. The specification of these controls is unique to this paper. These ten controls protect against the disclosure, re-identification, and harm risks associated with particular datasets. 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 and researcher or (b) 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 connections are blocked or (b) access to the Internet is allowed.

Output: (a) Iutput 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 or (b) researchers may view the research data.

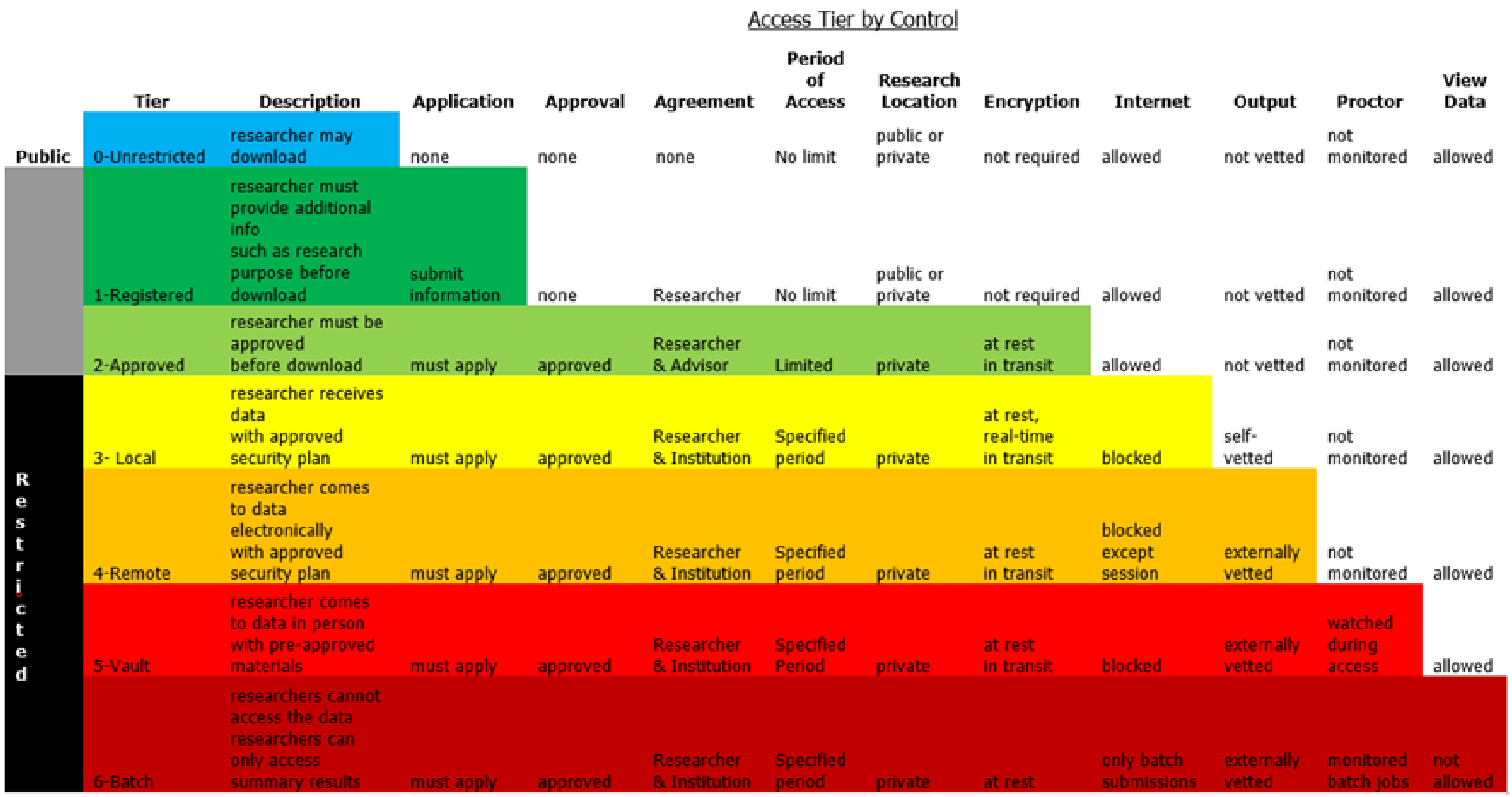
Figure 1 shows how the access tiers mesh with the controls. At each level, researchers must agree to and comply with all regulations. As risks increase, researcher agreement is not sufficient and technical configurations must prevent researchers from inadvertent disclosure. While researchers agree to follow all conditions, each tier adds a layer of security meant to ensure compliance. These extra security layers are, however, an impediment to research and should only be implemented when risks of re-identification and harm necessitate them.

Here are more details about each security control.

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, or confidentiality pledges. Furthermore, for data in tiers 3 and above, only researchers who serve as Principal Investigators (PIs) 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 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 alone or researcher and an institutional (or university or organizational) representative must sign the agreement to access the data. Tier 0-Unrestricted data do not typically require an agreement. Tier 1-Registered requires the researcher to agree to TOU, while tier 2-Approved also requires the agreement of an organizational official. For tier 3-Local and above, both the researcher and an institutional representative with authority to obligate the researcher's organization must agree to and sign the DUA.



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	none	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

Figure 1: Relationship between access tiers and controls.

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-Unrestricted and 1-Registered allow unlimited access. Agreements that require an institutional signature are always time-bound.

Research Location: is where the data will be viewed. A researcher's computer may be used to store the data or as a portal to a server where the data are stored. For tier 2-Approved and above, the client location must be private and specified. A private location prevents inadvertent viewing 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) require 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 a network connection. The requirement is usually met by the researcher agreeing to analyze the data on a standalone, non-networked computer. For tier 4-Remote, 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 network access. The purpose of blocking the Internet is to prevent researchers from inadvertently or otherwise copying files to unauthorized locations. An acceptable configuration implements a two-step process of copying files from 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 sub-sample sizes for regressions. Tier 3-Local and above require vetting, but all levels require compliance with rules about output. While tier 3-Local authorizes self-vetting, tier 4-Remote and above require that, in addition to the project team, trained personnel who are not part of the research project must review files before release. 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 accessing the data. For all tiers, researchers are not allowed to look up specific respondents in the data or to transcribe data points. For Tier 5-Vault and 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 microdata or only summary results. At Tier 6-Batch, researchers cannot view the microdata. While at all tiers, researchers agree to not attempt to re-identify or look up particular subject, at Tier 6, researchers are prevented from even accessing the microdata, so re-identification and lookups are impossible.

3. Seven Tiers, Five Safes, Three Access Categories

The seven tiers overlap with the Five Safes of Desai et al. [3] and the three access categories of Horton et al. [5] Horton's three categories are a combination of multiple tiers and form a similar hierarchy. The Five Safes overlap multiple tiers, and are a different conceptualization.

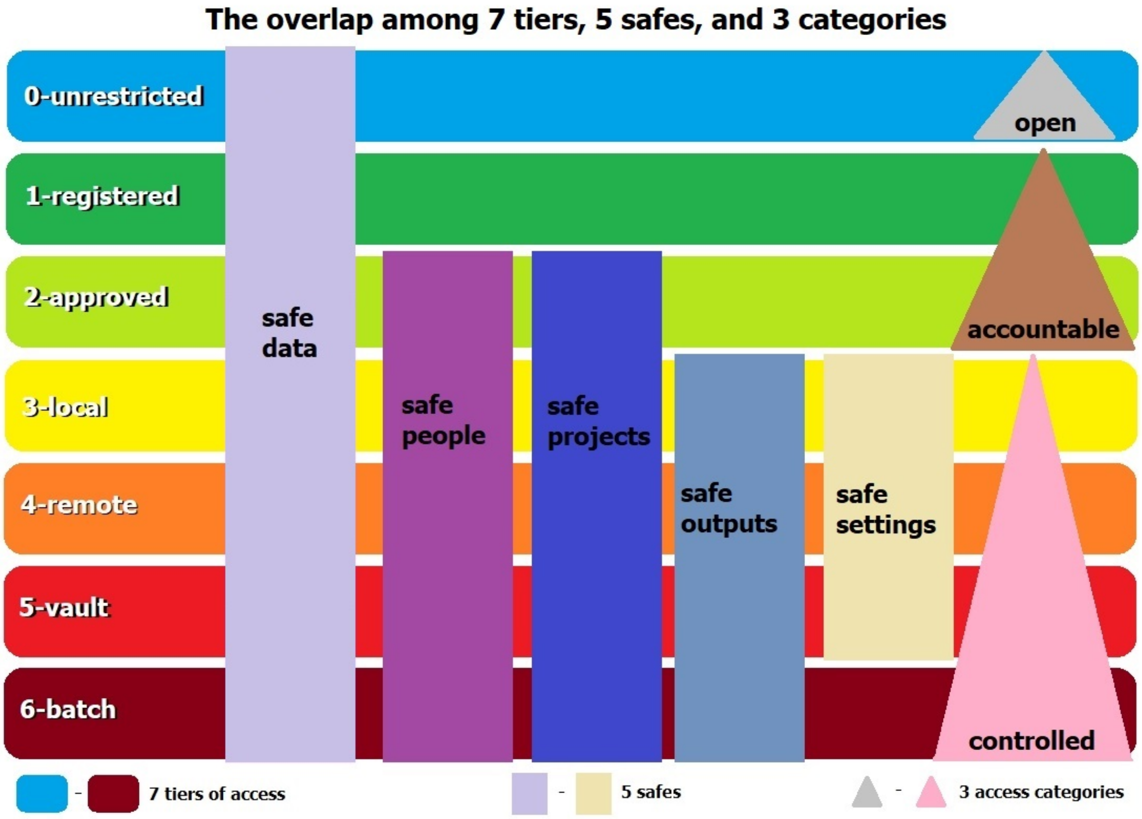


Figure 2: Overlaps among the seven tiers, five safes and three categories.

Safe people and projects are aspects of approved people and projects. Safe outputs correspond to tiers where output must be reviewed for compliance with disclosure protection rules. Figure 2 illustrates the overlaps.

Let us consider the implementation of the seven proposed tiers. As noted, all research data regardless of tier are for the calculation of summary measures only. Higher levels build on lower tiers by adding more security controls.

3.1. **0-Unrestricted.** Public-access research data are typically available for download from websites or via an Application Programming Interface (API). These data are available without restrictions on access. Disclosure and harm risks are negligible; nevertheless, the data are for research only. Unrestricted research data are also labeled public-use and open data. In many situations, public-access research data may be downloaded anonymously. The researcher who downloads the data can agree to the TOU without an institutional signature.

Example: Baby's First Years (BFY), New York City, New Orleans, Omaha, and the Twin Cities 2018-2021 [7].

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.

3.2. **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 contact information and (possibly) a research purpose; however, download of the data does not require approval. A researcher who downloads the data can agree to the TOU without an institutional signature.

Example: National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2018 [8] requires registration from anyone downloading the data.
Implementation: Registration system to collect information. Website and bandwidth to handle download demand.
Weakness: Researchers could provide inaccurate information.
Impediment to research: Researchers must provide information to access data.

3.3. **2-Approved.** Approved research data require registration and approval before download. The persons approving access 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 construed as sensitive. Because the data are available only upon approval, researchers may have to implement additional safeguards for these data, such as encryption. Researchers may only be allowed to access the data in a private setting. Researchers who download the data can agree to the TOU. In some cases, a second person, such as a department chair or graduate advisor, may need to supervise the research. The data collector in consultation with the data repository determines requirements.

Examples: The Panel Study of Income Dynamics (PSID) and Health and Retirement Study (HRS).
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.

3.4. **3-Local.** Research data in this tier are restricted; however, access to the data is at the researcher's own organization. These data have a higher risk of re-identification and harm if disclosure occurs. 3-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 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 requires IRB approval and confidentiality pledges from personnel who access them. 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: the NICHD Study of Early Child Care and Youth Development (SECCYD).

Implementation: Standalone (non-networked) computer in a locked private office. Some organizations may have an acceptable server set up.

Weakness: The research data are not under the control of the repository. Unauthorized access is possible.

Impediment to research: It may be difficult to collaborate with a research team. Organizations may be reluctant to permit a non-networked computer. Researchers may not have funds to buy a computer that is dedicated to a single project.

3.5. **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 may have higher re-identification and harm risk. In some cases, the 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 a local computer so that the repository retains control over access to the data. Researchers must review their output for compliance with disclosure protection rules; trained repository staff must also vet the output. Only files that meet disclosure protection requirements are released from 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) data.

Implementation: Terminal server or Virtual Desktop Infrastructure (VDI) that prevents files from being copied from the server or virtual instance. The Inter-University Consortium for Political and Social Research (ICPSR), NORC, and Survey Research Center at the Institute for Social Research (University of Michigan) 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 computational power of virtual machines may not be sufficient for some research.

3.6. **5-Vault.** Vault protocols add a proctor or observer to the security requirements. A "vault" is a locked room where data can be accessed only 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. The data typically have even higher re-identification and harm risks.

Example: Videos, which almost always pose high disclosure risk. As with data in 4-Remote, all output and files are reviewed before release.

Implementation: Locked room with proctor. The Federal Statistical RDCs have implemented this level of security. Some research centers have vault rooms for accessing restricted data.

Weakness: Researchers could still look up individual records.

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.

3.7. 6-Batch.

This tier is for research data with the highest risks, and provides the maximum protection since researchers are unable to view the microdata. 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, cross-tabs or other analyses 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 they released to the researcher. Accessing these data requires an application and approval as well as institutional agreement. While this tier does not allow researchers to view the microdata, this level has one advantage over 5-Vault in that it does not require travel.

Examples: *LISSY* at the Cross-National Data Center in Luxembourg; the retired *ANDRE* system at the National Center for Health Statistics (NCHS).

Implementation: Batch system. A server with (possibly 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 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 specifications show how each tier adds controls to ensure compliance. For research data with human subjects, tiered access is essential since data vary in their risks of re-identification and harm.

4. 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, as well as examples of datasets that fall within each tier. With these seven tiers of access, repositories can meet the needs of researchers while still providing appropriate protection for research data. The tiered approach enables repositories to impose sufficient security controls without creating unnecessary impediments to research.

References

[1] M. Crosas. CIO Review: Cloud Dataverse: A Data Repository Platform for the Cloud. URL: https://openstack.cioreview.com/cxoinsight/cloud-dataverse-a-data-repository-platform-for-the-cloud-nid-24199-cid-120.html.

[2] Data Protection Commission (Ireland). Guidance for controllers on data security, 2020. URL: https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Data_Security_Guidance_Feb20.pdf.

[3] T. Desai, F. Ritchie, and R. Welpton. Five safes: designing data access for research. Technical report, University of the West of England, Bristol, UK, 2016. Economics Working Paper Series 1601. URL: https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf.

[4] Federal Committee on Statistical Methodology Confidentiality and Data Access Committee. Restricted access procedures, 2002. URL: https://nces.ed.gov/FCSM/pdf/CDAC_RAP.pdf.

[5] L. Horton, A. Perry, and L. Bishop. Open where possible, closed if necessary: reforming access categories for social science data archives. In *International Digital Curation Conference 2020*, Dublin, Ireland, 2020. https://doi.org/https://doi.org/10.5281/zenodo.3670943.

[6] S. K. Kinney, A. F. Karr, and J. F. Gonzalez Jr. Data confidentiality: The next five years summary and guide to papers. *Journal of Privacy and Confidentiality*, 1(1), 2010. https://doi.org/https://doi.org/10.29012/jpc.v1i2.569.

[7] K. A. Magnuson, K. Noble, G. J. Duncan, N. A. Fox, L. A. Gennetian, H. Yoshikawa, and S. Halpern-Meekin. Baby's First Years (BFY), New York City, New Orleans, Omaha, and Twin Cities. Inter-University Consortium for Political and Social Research 2020-11-16. https://doi.org/https://doi.org/10.3886/ICPSR37871.v2.

[8] K. Mullan and J. R. Udry. National longitudinal study of adolescent to adult health (add health), 1994-2018, 2022. Carolina Population Center, University of North Carolina at Chapel Hill and Inter-University Consortium for Political and Social Research, distributors. https://doi.org/https://doi.org/10.3886/ICPSR21600.v24.

[9] J. P. Reiter and S. K. Kinney. Commentary: Sharing confidential data for research purposes: A primer. *Epidemiology*, 22(5):632–635, 2011. https://doi.org/https://doi.org/10.1097/EDE.0b013e318225c44b.

[10] A. Slavkovic, S. K. Kinney, and A. F. Karr. O privacy, where art thou? *Chance*, 24(4):41–45, 2013. https://doi.org/https://doi.org/10.1080/09332480.2011.10739886.

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