Restricting data's use: A spectrum of concerns in need of flexible approaches[1]

Dharma Akmon[2], Susan Jekielek[3]

Abstract

As researchers consider making their data available to others, they are concerned with the responsible use of data. As a result, they often seek to place restrictions on secondary use. The Research Connections archive at ICPSR makes available the datasets of dozens of studies related to childcare and early education. Of the 103 studies archived to date, 20 have some restrictions on access. While ICPSR's data access systems were designed primarily to accommodate public use data (i.e. data without disclosure concerns) and potentially disclosive data, our interactions with depositors reveal a more nuanced range of the needs for restricting use. Some data present a relatively low risk of threatening participants' confidentiality, yet the data producers still want to monitor who is accessing the data and how they plan to use them. Other studies contain data with such a high risk of disclosure that their use must be restricted to a virtual data enclave. Still other studies rest on agreements with participants that require continuing oversight of secondary use by data producers, funders, and participants. This paper describes data producers' range of needs to restrict data access and discusses how systems can better accommodate these needs.

Keywords

data archives, restricted access systems, privacy, confidentiality

Introduction

Responsible stewardship of data requires the ability to restrict access when data could identify individuals and potentially cause harm through their disclosure. At the same time, access restrictions by definition limit data's use, so data archives must also apply restrictions judiciously, ensuring restrictions match the level of disclosure risk. ICPSR[4], a membership-based social science data archive, currently offers three main dissemination options for sensitive data: secure download, virtual data enclave (VDE), and physical data enclave. Each of these dissemination options imposes a stringent application process, but differ in where the data are accessed and what is required to access the data. ICPSR's stepwise series of dissemination options have been devised to serve sensitive data that vary in their probability of disclosing individuals and the potential severity of harm were individuals' information disclosed.

In this paper, we discuss ICPSR's options for restricting access to data using three examples from one of its topical archives: Childcare and Early Education Research Connections[5]. These examples demonstrate some of the reasons for restricting access to data. In doing so, they also highlight the ways in which the need for restricting use does not always align with the systems and tools we have implemented for accessing the data. After describing the examples, we discuss design implications for

Akmon, Dharma and Jekielek, Susan (2019) Restricting data's use: A spectrum of concerns in need of flexible approaches, IASSIST Quarterly 43(3), pp. 1-7. DOI: https://doi.org/10.29173/iq941

restricted access systems that will better serve the needs of the data, researchers, and study participants.

ICPSR and Restricted Access Options

ICPSR, founded in 1962, is a consortium based at the University of Michigan of over 760 member institutions around the world to archive and make social and behavioral science data available to researchers. ICPSR holds over 10,000 data collections, approximately 1,670 of which have at least one dataset with restrictions on its use. In restricting access to data, ICPSR aims to protect the confidentiality of study participants and ensure that the benefits of research outweigh the potential harms to individuals. Information such as criminal activity, antisocial activity, and medical conditions could cause harm were it associated to particular individuals, who could be ascertained through direct or indirect identifiers. Direct identifiers include name, phone number, social security number, and location, while indirect identifiers include information that can be used to identify a subject when combined with other information (for example, gender, birth date, geographic indicator and other descriptors). The only way to ensure 100% confidentiality protection is by blocking access to data. As data stewards, we must balance the tradeoff between the value of making the data available for others to use in new research and the responsibility to protect the confidentiality of study participants.

The vast majority of data archived with ICPSR are public use files: these are data that ICPSR, often along with the data producers, have assessed to present very little, if any, risk of harm and/or probability of disclosure. A secondary user accesses these data through a simple web download: she logs into her ICPSR MyData account, agrees to terms of use by checking a box, and after doing so, the data download immediately to her local machine. For sensitive data, ICPSR staff use several approaches to maintain confidentiality: they modify data to reduce the chance of reidentification (e.g. through masking data); physically isolate the data and use secure technologies to provide access; train researchers in the responsible use of the data; and require researchers to agree to particular guidelines of use through restricted use agreements. Typically, the agreement includes a responsible use statement, a research plan, Institutional Review Board (IRB) sign-off for the plan, a data protection plan, behavior rules (e.g. the researcher will not attempt to identify individuals and will not share the data with others), a security pledge, and an institutional signature on the agreement to use the data.

Data with moderate risk of harm and probability of disclosure (in other words, data with some risk of reidentification, disclosure, and non-trivial risk to study subjects), are generally offered via ICPSR secure download. The secure download option requires a researcher to submit an IRB-approved research proposal; a data use agreement signed by her institution; an agreement that the data will only be used within a very particular computing configuration (for example, on a stand-alone machine that is not connected to the Internet); and an affidavit of the data's destruction at the conclusion of the use period (generally one year, but the researcher can apply for extensions to the agreement). Once approved, researchers receive an encrypted file via email, which they download to the secure location specified in the application.

For higher risk data—that is, data that might be reidentified and also cover sensitive topics where disclosure could harm the study participants—ICPSR offers both a virtual and physical data enclave.

Each of these enclaves require a similar set of application materials as for secure download, however, the data can only be accessed and analyzed from within a highly restricted environment. In the case of the virtual data enclave (VDE), researchers access the data through a virtual machine they launch from their own computer but that operates on a remote server. The virtual machine is completely isolated from the user's physical computer, restricting the user from downloading files or parts of files to their physical computer. The virtual machine is also restricted in its external access, preventing users from emailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally. To receive final output from the VDE, the data must be vetted by ICPSR staff. Furthermore, ICPSR has the capability to shut off access to the data when the terms of the data use agreement end and in the rare case of a user violating the terms of the data use agreement.

Data that are deposited in the physical enclave can only be accessed on site in Ann Arbor, Michigan. The machines in the physical enclave are not connected to the Internet, and an ICPSR staff member is present at all times when a researcher is using the enclave. ICPSR staff conduct a disclosure review of all files that the researcher wants to use after leaving the enclave.

Three Examples of Restricted-Use Data

ICPSR provides data dissemination for more than 20 federal and non-governmental sponsors via topical archives on topics such as addiction and HIV, aging, arts and culture, and criminal justice. ICPSR's Childcare and Early Education Research Connections archive (hereby referred to as "Research Connections") is funded through the Office of Planning, Research and Evaluation, Administration for Children and Families (OPRE), U.S. Department of Health and Human Services and curates and provides access to over 300 studies. As an archive at ICPSR, Research Connections has at its disposal the restricted-use access options described above and works closely with both the sponsor and data producers to ensure that confidentiality is maintained while facilitating the broadest use of data possible.

Working closely with study stakeholders has given the Research Connections project team the opportunity to understand the myriad concerns at play when making sensitive data available to secondary users. For the purposes of this paper, we discuss three studies from Research Connections that demonstrate the nuanced needs of restricting access to data.

The Head Start Family and Child Experiences Survey

The first study we discuss is the Head Start Family and Child Experiences Survey[6], also known as "FACES." The FACES study provides longitudinal data on a periodic basis on the characteristics, experiences, and outcomes of Head Start children and families as well as the characteristics of the Head Start programs that serve them. FACES also provides information on the relationship among family and program characteristics and outcomes. Several cohorts of FACES have been fielded since 1997, and, through Research Connections, ICPSR has curated and provided access to data collected from six cohorts of this study.

As study sponsors and data producers work with Research Connections to make their data available to secondary users, they are concerned not only with the potential disclosure of individual study participants, but also with the disclosure of particular centers (e.g. a specific Head Start center), since such information could potentially cause harm to the center and the families it serves. For that reason, ICPSR works to ensure that both personal and center characteristics are non-disclosive. With all waves of FACES, the direct and indirect identifiers have been removed from the data made available through Research Connections, leaving virtually no chance that individuals or centers could be reidentified. These protective actions might suggest treating this study as a public-use study, and that simple download, where nothing more than logging in and agreeing to ICPSR's standard terms of use is required, would be the most appropriate dissemination method. Yet, because of the moderate risk of harm to the centers should they be reidentified, the sponsor and data provider were keen to place additional requirements on researchers that want to access and use the data. They wanted to both lightly screen potential researchers to ensure that they are using the data for legitimate research or public policy purposes and create a means of addressing data misuse should it occur. ICPSR's simple download currently provides no means to meet those goals, but the standard secure download does. However, no one involved felt the disclosure of the data was sufficiently risky to warrant additional barriers to researchers for using the data, including IRB sign-off, a signed agreement from the researcher's institution, and strict technology prerequisites.

Data collected from the earliest FACES cohorts predated ICPSR's online restricted data application system, and data access was administered via a paper application form. The advent of ICPSR's online restricted data access system brought some challenges for ICPSR to consider for the dissemination of FACES. Because simple download did not satisfy the dissemination needs of the study, and the new on-line application system placed more barriers than were necessary to responsibly disseminate the data, we developed a hybrid approach. For FACES, we continue to gather the required researcher information through the original paper form that researchers download from the study's homepage, and then—once we approve the researcher's access—we deliver those data via a download link emailed to the user. Once the researcher gets the data, the data are subject to the same rules as public-use data: the researcher is not to redistribute the data or attempt to identify individuals or organizations and must properly cite the data in any publications they make from using the data, but they are not required to destroy the data upon completion of their work with them.

The Head Start Impact Study

The second example demonstrating the need for flexible technical approaches comes from the Head Start Impact Study[7] (HSIS), which is a nationally representative sample of Head Start programs and over 4,500 children. As with FACES, the disclosure of centers presents moderate risk of harm to the Head Start centers included in the study. While all of the study's direct identifiers have been removed, some indirect identifiers are available to enhance the analytic utility of the data. The concerns with disclosure were, therefore, greater than they were for FACES, and the data producers and sponsor agreed that the restrictions associated with the secure download option were most appropriate. However, they were highly concerned with making the most sensitive of the study's files available through secure download. The Center Analysis File contains a compilation of publically available population and household characteristics data for each center's local community at the county and census tract level. Even though centers are not directly disclosed in the file, because the information in the file is unique for communities, the centers could conceivably be identified through triangulation with public data sources. As a result, the data producers and study sponsor had serious concerns with allowing researchers to download the file, even with the strict technology requirements of secure download. In fact, the data producers were only willing to make this file available through the VDE, where researchers must confine their work with and analysis of the data. Only vetted, final output can be removed from the VDE, further reducing the disclosure risk for this file.

In some ways, the needs of HSIS fit very well with ICPSR's options for restricting access. However, the inclusion of a single file requiring VDE dissemination within a study where every other file could be offered via secure download somewhat challenged technical systems that do not easily allow us to specify a different means of dissemination for a single file within a restricted use study. Therefore, we had to create what essentially became two different studies at ICPSR: 1) the Head Start Impact Study[8], where files could be applied for and accessed via secure download and 2) the Head Start Impact Study with Center Analysis File[9], where researchers could apply and access all of the studies files within a VDE. This means that whether or not a researcher needs to work with the Center Analysis File dictates which version of the study she should apply to access.

American Indian/Alaska Native Head Start Family and Child Experiences Survey

The last example we discuss to better understand restricted data access needs is the American Indian/Alaska Native Head Start Family and Child Experiences Survey[10] (AI/AN FACES). AI/AN FACES stands as the first ever national study on AI/AN Head Start Children and families. As with the FACES study discussed earlier, there is a moderate risk of harm to centers if disclosed. The risk of disclosure is somewhat higher due to narrow focus of the study on a particular population, however direct identifiers and other indirect identifiers were removed. What makes this study unique and adds nuance to how we should think about restricting access is the study's focus on a specific community and the associated agreements that the sponsor and data producer have with the tribal communities that participated. Specifically, researchers' permissions to use of the data depends on their commitment to use the data with consideration for the tribal centers in which the data were gathered. This includes both training and a researcher review process that is far outside the scope of ICPSR's internal restricted use application review and requires reviewers with expertise in working with tribal communities.

The risk of disclosure and harm associated with disseminating AI/AN FACES best fit with the requirements and application process for ICPSR's secure download dissemination option. In addition, for the aforementioned reasons, we had to meet the study's need for third-party review of applications, which none of ICPSR's restricted use options were designed to do. We created a workflow that leveraged both the study's homepage and secure download application system, resulting in what we might think of as "secure download plus." Researchers must follow two separate application processes that serve different functions and come together in the final submission package to ICPSR. On the AI/AN FACES homepage, interested researchers access the *AI/AN FACES Application Guide*[11], which includes instructions and guidelines for submitting an application package to the third-party panel (the AI/AN FACES Data Committee) as well as links to required trainings to use the data.

All research plans are reviewed by the external AI/AN FACES Data Committee, which is comprised of individuals with expertise in conducting research with tribal communities as well as representatives from Native American Head Start programs or tribal community representatives. The committee reviews each research plan to evaluate whether the research plan demonstrates a sound understanding of the study design and sample and the proposed research questions are able to be answered by the data; assess the plan for disseminating findings; and evaluate the expertise and experience of the research team in working with tribal communities.

At the same time, researchers follow the ICPSR restricted use application process. In the end, the researcher submits an application to ICPSR that includes an IRB approval/exemption notification letter, the signed AI/AN FACES Data Committee[12] notification letter, a signed acknowledgement that the researcher has read *Best Practices for Working with AI/AN FACES Data*[13], and a signed data use agreement. After everything has been approved, researchers receive the encrypted files via email, as they would with ICPSR's standard secure download process.

Conclusions

The three studies we described reveal nuance in the dissemination needs of sensitive data. Specifically, these examples show that flexibility is needed around the following three key areas: the information collected from would-be users as a prerequisite to data delivery; the differing degrees of sensitivity of individual files within the same study; and who to include in the review and approval of applicants to use the restricted data.

Like FACES, some studies represent enough disclosure risk to warrant collecting additional information about applicants and screening them prior to disseminating the data. As our workaround shows, building in the capability to specify for particular studies which information about a researcher must be collected (e.g. full name, affiliation, research questions, contact information) *without* the more stringent requirements of IRB sign-off, institutional legal agreements, and strict technology setups would help appropriately protect study participants without placing unnecessary obstacles to accessing and using the data. The FACES study requires an ICPSR staff member to review and approve applicants, but we can also imagine scenarios where it is sufficient to collect additional information about secondary users (i.e. not null) without requiring staff review and approve the researcher's access to the data, suggesting another area where flexibility is needed.

As the Head Start Impact Study demonstrates, individual datasets within a study can represent different levels of disclosure risk, and so our systems also need to allow us to specify dissemination restrictions at the file level. While ICPSR's current systems allow for a mix of restricted-use and public-use files within the same study, they do not easily allow for the restricted files within the same study to have varied dissemination methods. Customizable settings that allow staff to specify different access requirements for different restricted-use files promise to improve the user experience for researchers who currently must determine which version of the study they need and then are directed to completely different application systems (one for secure download; another for virtual data enclave) depending on the files of interest.

Our third example demonstrates that some studies require more specialized vetting of users than can be provided by repository staff on their own, particularly when consent agreements define a particular review process as a prerequisite to data access. Studies such as AI/AN FACES demand not only IRB sign-off, institutional agreements, and strict technology configurations, but also a commitment that all interested researchers complete specialized training and have their research plans approved by a panel of individuals with expertise specific to the study. ICPSR's current system met this need with an existing application portal that allows for the inclusion of additional documents in the submitted application package. Interested researchers follow two application processes—ICPSR and the AI/AN FACES panel—that come together in the end with a single application package to ICPSR.

Finally, all three examples demonstrate the importance of building in documentation and review of individual case studies to support continuous system improvement. While ICPSR's systems are flexible enough to provide solutions for each of the cases described, the details of each case can be used to inform future system development that facilitates researcher access to and analysis of secondary data while maximizing protection of data. In this way, data repositories more effectively balance the need to protect study participant confidentiality with facilitating the broadest data reuse possible and bolster their credibility as responsible stewards of research data.

[1] This paper was originally presented at IASSIST 2018, Montreal, Canada.

[2] Dharma Akmon is Assistant Research Scientist and Director of Project Management and User Support at the Inter-university Consortium for Political and Social Research ICPSR and can be reached by email: dharmrae@umich.edu

[3] Susan Jekielek is Assistant Research Scientist and Director, Education Archives at the Inter-university Consortium for Political and Social Research and can be reached by email: jekielek@umich.edu

[4] https://www.icpsr.umich.edu/icpsrweb/

[5] https://www.researchconnections.org/childcare/welcome

[6] https://www.researchconnections.org/childcare/series/236

[7] https://www.researchconnections.org/childcare/studies/36968

[8] https://www.researchconnections.org/childcare/studies/29462

[9] https://www.researchconnections.org/childcare/studies/36968

[10] https://www.researchconnections.org/childcare/studies/36804

[11] https://www.researchconnections.org/files/childcare/pdf/AIAN_ICPSR_Guide.pdf

[12] The AI/AN FACES 2015 Data Committee is comprised of individuals with expertise in conducting research with tribal communities and representatives from Region XI AI/AN Head Start programs.

[13] https://www.researchconnections.org/files/childcare/pdf/AIAN_FACES_Best_Practices_for_Researchers.pdf