Version 3-2023

INSTITUTIONAL REVIEW BOARD

Data Management Handbook for Human Subjects Research

An educational resource intended to help researchers who want to conduct research with human participants construct an effective data management plan as part of their research proposals.

by Alena Filip
Human Protections Analyst
Institutional Review Board (IRB)
Office of Research

Table of Contents

List of Tables

List of Figures

List of Acronyms

COC	Certificate of Confidentiality
CSU	California State University
DUA	Data Use Agreement
FERPA	Family Educational Rights and Privacy Act
HIPAA	Health Insurance Portability and Accountability Act
OHRP	Office of Human Research Protections
ISO	Information Security Officer
HSR	Human Subjects Research
IRB	Institutional Review Board
NIST	National Institute of Standards and Technology
PI	Principal Investigator
PII	Personally Identifiable Information
PPRA	Protection of Pupil Rights Amendment
RSCA	Research, Scholarship, and Creative Activities
SBER	Social, Behavioral, and Education Research
SJSU	San José State University
SSN	Social Security Number

List of Supplemental Documents

These documents, provided by the Office of Research on behalf of the IRB, are designed to work in conjunction with this handbook:

- Data Management Checklist
  A 2-page checklist based on this handbook that summarizes the most important points.

- Data Management Excel Template
  A spreadsheet template that investigators can use to manage their use of data throughout the lifecycle of the research.

Introduction

A comprehensive data management plan entails understanding the content of the data and how it will be handled during its life cycle from collection, processing, dissemination, to disposal or archiving. Data management involves knowledge of both information security and privacy concepts. This handbook is intended to help investigators conducting human subjects research (HSR) construct a data management plan and to provide informational resources about data management. We focus primarily on issues surrounding social, behavioral, and education research (SBER) as this is the most common area of research involving human subjects at San José State University.[1] Reference to applicable SJSU or CSU policies are provided in the relevant sections under the chapter "Elements of an HSR Data Management Plan." This handbook serves as a guide to data management issues and applicable regulations. It is not meant to give legal advice. As well, investigators should not confuse the elements of data management listed in this handbook with other elements that may be required as part of data management plans included in a grant application submission. Make sure to check with the sponsor of the research for funding-specific requirements.

In order to adequately assess risk to human subjects, Institutional Review Boards (IRBs) must understand the overall data management plan for a specific research project. Though not all elements may need to be communicated to the IRB, investigators are encouraged to use this handbook to become better informed about their responsibilities as data stewards.

Data Management Glossary

IRB members, in their review of research applications, frequently point out the confusion over commonly used terms such as "anonymous", "confidential", and "private." Casual use of terminology by investigators not only hinders effective compliance review, but it also undermines consistency in communication with the public about research activities. Investigators who are not mindful of their use of data management terms also risk being found non-compliant with existing policies and regulations if they are not able to deliver the protections promised to research participants who have volunteered to disclose personal information about themselves.

Data involving human subjects can be thought of as existing on a spectrum of identifiability (Emam, 2010; Garfinkel, 2015, p. 5). Figure 1 shows the degree of identifiability for some of the different terms outlined in our glossary.[2] The definitions that follow are critical to understanding effective data management practices. Principal investigators (PIs) are responsible for appropriately incorporating these data management terms into their IRB applications, into their training of research personnel and, most importantly, into their communications with research participants.

[1] For more technical guidance on statistical techniques for managing data refer to George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice* (2011).

[2] For a slightly different breakdown of the terms outlined in this handbook, refer to the useful infographic A Visual Guide to Practical Data De-Identification (The Future of Privacy Forum, 2017).

Potentially Public Data Types

Anonymized: Data cannot be associated with an individual; re-identification is not possible.

De-identified: Data cannot be readily linked to individuals because direct and indirect identifiers are removed before the data are released.

Aggregate: Data can be ambiguously linked to several individuals but not to a specific individual.

Masked: Data can be linked to a specific individual but are only reported after masking techniques have been applied; direct and potentially sensitive indirect identifiers are replaced before data are reported. **Pseudonymization** is an example of a simple masking technique. Other, more advanced masking techniques may utilize statistical manipulation.

Corresponds to Level 3 data in the SJSU data classification scheme.

Potentially Protected Data Types

Personally Identifying Information (PII): Data are linked to a specific individual and include direct and indirect identifiers that are typically accessed for internal use but are not typically disseminated to the public. Confidentiality protections typically apply to the data when it is used for human subjects research, though in other contexts PII may not be protected.
Examples: name, partial date of birth, non-directory student information

Corresponds to Level 2 data in the SJSU data classification scheme.

Protected and Private Data Types

PII Subject to Additional Protections: Data are linked to a specific individual and include direct identifiers that are subject to both confidentiality and privacy protections, which can include laws and internal administrative policies. Internal access is also limited on a "need to know" basis determined by a person's role and job function within an organization.
Examples: social security number, medical record number and records, biometric information

Corresponds to Level 1 data in the SJSU data classification scheme.

Proper de-identification moves data up on the spectrum

Figure 1. Spectrum of Identifiability.

Aggregate Data: Generalization of data to broader categories that cannot be readily manipulated to re-identify specific research subjects. Small sample sizes tend to reduce the possibility of aggregating data.

Anonymized vs. De-identified Data: The term *anonymized* refers to data from which identifying information about research participants, both direct and indirect, is permanently and completely removed. Anonymized data can no longer be associated with an individual in any manner. Once this data is stripped of personally identifying elements, those elements can never be re-associated with the data or the individual by anyone.

De-identified is another term that is often used synonymously with anonymized. However, in many guidance[3], the two terms are characterized by a subtle difference: A de-identified data set still holds the possibility for re-identification, whereas an anonymized data set does not. De-identification is a general term for the process of removing the association between a set of identifying information and the individual who provided it, while preserving the utility of the data for other purposes (Garfinkel, 2015, p. 8). The techniques used for de-identifying data – which amount to transforming the data by removing, suppressing, blurring, or masking certain elements – are typically applied prior to sharing or releasing a dataset as part of research findings, as well as for long term storage of institutional data. De-identification minimizes privacy risks to the data subjects but does not eliminate risk.

For SJSU IRB purposes, we use the term *de-identified* to refer to the status of the data – that the identity of individual research participants cannot be readily deduced once any identifiers have been removed but "with the understanding that sometimes de-identified information can be re-identified and sometimes it cannot" (Garfinkel, 2015, p. 3).

We use the term ***anonymous*** to denote the relationship between the research team and the research participants – that the identity of the research participants is not known to the research team because the participants are randomly recruited, there is no direct interaction between the research team and the participants, no personally identifiable information is collected or recorded, and no recording or tracking tools are used with the data instruments. A feasible example where all of these criteria are met is an online survey that does not collect any direct or indirect identifiers, with disabled tracking features, advertised in a public forum.

We use the term *anonymized* to refer to data when they are irrevocably de-identified with no possibility of re-identification. Unfortunately, the prolific overuse of terms related to anonymity is seldom justified; the terms *anonymous* and *anonymized* seem to be increasingly theoretical rather than practical descriptions and they should be used with informed caution.

Investigators who obtain secondary data that have already been de-identified from a third party are not required to have their research evaluated by the IRB, as the activity is considered to not involve human subjects (45 C.F.R. § 46.102(e)(1), 2018). On the other hand, investigators who will collect, retrieve, receive, or use identifying data about human subjects with plans to de-identify it must submit

[3] See, for example, Educause, "Guidelines for Data De-identification," (2015); The Future of Privacy Forum (2017).

an application to the IRB, as must investigators who plan on administering a completely anonymous survey that they have designed.

Techniques for de-identification are shared in this document, but it is worth keeping in mind that there is no absolute method for protecting the direct and indirect identifiers of research participants from re-identification. Even data that are considered to be de-identified can potentially be recovered or reconstructed with techniques and technology that seem to be evolving faster than security measures can keep up with (Nelson, 2015, p. 19). It's been said that "data can either be useful or perfectly anonymous but never both" (Ohm, 2010, p. 1704) and, "as long as any utility remains in the data derived from personal information, there also exists the possibility, however remote, that some information might be linked back to the original individuals on whom the data are based" (Garfinkel, 2015, p. 1). Researchers should not only be mindful of appropriately balancing the risk of re-identification with the potential value of the data they wish to collect, but they are responsible for applying strategies for reducing the risk to acceptable levels and understanding the limitations of de-identification. As described in published standards, "there are no widely accepted standards for testing the effectiveness of a de-identification process or gauging the utility lost as a result of de-identification" (Garfinkel, 2015, p. 39).

Confidential Data: The status of data that are protected from unauthorized disclosure to unauthorized individuals or entities. When it comes to data collected from human subjects, confidential data is not anonymous data by definition; research participants are assured that any potentially identifying information they share will not be disclosed outside of the research context if confidentially is promised by the research team. SJSU's Risk Assessment Program Standard (2015) also considers misuse of data beyond the scope of duties by those authorized to use it to be a violation of confidentiality (p. 6). For researchers, this means that primary data collected from human subjects should only be used for the purposes stated on the research consent document. Secondary data may also be designated as confidential by institutional policies and/or by state and federal regulations; use of such data for research must be approved by an IRB review process, which will determine whether consent from the original data subjects is required. In addition, there may be other institutional permissions that are needed to access certain data. Confidential data, as a subset of sensitive data, is generally considered to have more restrictions around use and disclosure.

Data Sanitization: The process of removing information from storage media (including paper and digital records) such that data readability and recovery is not possible. This term can also refer to the process of disguising sensitive information in information technology resources by overwriting it with realistic looking, but false, data.

Direct Identifiers: Data elements that are unique to a single individual and can be used to identify him/her (e.g., name, social security number, student ID, email address, photograph, HIPAA privacy rule 18 identifiers, biometric data).

Indirect Identifiers: Data elements that by themselves do not identify a specific individual but can be combined or connected with other information to identify the individual (e.g., racial or ethnic identity,

gender, date or place of birth, location information, HIPAA limited data set, social media handle/usernames).[4]

Indirect identifiers, also sometimes called "quasi-identifiers," pose challenges to de-identification of data because they "generally convey some sort of information that might be important for a later analysis and removing them may damage the utility of the dataset. As such, they require careful consideration to balance the risk of re-identification with the utility gained by their inclusion" (Garfinkel, 2015, p. 20).

Personally Identifiable Information (PII): Very broadly, PII is any data that could potentially identify a specific individual or that can be used to distinguish one person from another (e.g., name, social security number, student ID). However, there is no general, over-arching, agreed upon definition, nor can a definitive list of what constitutes PII be provided. Part of the challenge in defining PII is that the term maybe used differently according to various laws, regulations, and guidance documents. Another challenge is that the meaning of PII may differ based on context – how the data are collected, connected, and reported, and the availability of other data sets that can be compared with and tied to research data. Some guidelines include, as part of the definition of PII, the ability to trace an individual's identity through linkages with other information that may not be part of the original data set.[5]

To complicate matters, different regulations may have different rules about what constitutes identifying information. To illustrate this, see Table 1, which compares the federal Policy for the Protection of Human Research Subjects with the HIPAA Privacy Rule for the protection of medical records.

Table 1 *Differences in the Definition of Identifiable Information between the Federal Policy for the Protection of Human Research Subjects and the HIPAA Privacy Rule*

Federal Policy for the Protection of Human Research Subjects	**HIPAA Privacy Rule**
States that the key mapping the coded data to identifiers must be destroyed for the data to be considered anonymous	Allows you to create and keep (and protect) a key that maps the coded data to identifiers
Allows for the code used for data elements to be derived from identifying information	Does not allow the code used for data elements to be derived from identifying information
States that anonymized data should not contain any identifiable private information but excludes ZIP codes and dates from the definition of "identifiable"	Lists 18 specific data types to be considered identifying info, including zip codes and dates

Adapted from text in "De-Identification of Clinical Trials Data Demystified," by Shostak (2006).

[4] It's important to keep in mind just how few indirect identifiers are needed to re-construct identifying information. A combined date of birth, zip code, and gender has been shown to uniquely identify up to 87% of the U.S. population (Sweeney, 2000).

[5] See, for example, the Institute of Education Sciences, National Center for Education Statistics, SLDS Technical Brief, "Data Stewardship," (2010), p. 2.

Pseudonymization / Coding: A type of data de-identification technique that removes identifiers associated with the subject and replaces them with a fake name or code in order to associate a particular set of characteristics or data records relating to a specific subject. The term *coding* is sometimes used interchangeably with pseudonymization.

Pseudonymization/Coding is considered to be a useful security measure because it reduces the linkability of research-specific data with other data sets (Garfinkel, 2015, p. 17). However, data transformed by this technique are not considered to be fully anonymized so long as a link (also called a key) between the identity of the subject and the code is preserved or can be re-generated. Also, if a code is retained over a long period or is linked to an increasing amount of information about an individual, there is an increased risk to the privacy of the original data subject (Ibid.).

Many regulations, including FERPA and HIPAA, do not allow the pseudonym or code to be derived from personally identifying information. [6] Some effective methods for coding include using a random number generator or using a sufficiently generic fake name (rather than a nickname). PIs are responsible for limiting the amount of people who have knowledge of the method used to generate the code and the people who have access to the key.

Private Data: Highly restricted information that is not typically shared with anyone or is only shared with a select and limited number of people who have a legitimate need to access the data. The purpose of designating data as "private" is to mitigate severe potential risks to individuals and institutions. In human subjects research, privacy is the concept that individuals have the freedom to not be observed and have autonomy over how their person or personal data are accessed. Some privacy concepts, such as the HIPAA Privacy Rule for the protection of personal health information and medical records in the U.S., have been codified into law. In other cases, individuals may have the expectation of privacy even when there is no specific law that protects them. For example, individuals have the expectation that when they use a public restroom, their activities will not be monitored.[7]

Additionally, "private data" can refer to the highly restricted status of security-related information (e.g., passwords, encryption keys, and other forms of authentication) that a research team puts in place for protecting the HSR data in their possession.

Private data, as a subset of confidential data, is generally considered to have the highest restrictions around use and disclosure.

Protected Data: A term used in this handbook to denote the status of Level 1 and Level 2 data, as described by the SJSU Information Classification and Handling Standard (2019).

---

[6] FERPA has additional requirements for the use of secondary coded data in research, including that the data can only be used for educational research purposes, that the party receiving the data is not allowed any access to the information about how the descriptor is generated and assigned, and that the code cannot be used to identify the student or to match the information from education records with data from any other source (as described in U.S. Department of Education, Privacy Technical Assistance Center, "Data De-identification," 2012, p. 5).

[7] This expectation of privacy has, unfortunately, not been observed by researchers in the past. See the Tearoom Trade study, conducted in 1970, as an example (Wikipedia, Accessed April 10, 2019).

Sensitive Data: An umbrella term used for any data that may need to be protected from unwarranted disclosure but that is not always subject to formal policies or laws. The CSU Information Security Manual defines sensitive data as "information which must be protected due to proprietary, ethical, contractual or privacy considerations" (Section 8065.S02, 2011, p. 2). The sensitivity of data depends on the context of use, the availability of PII, the importance of the information to the data subject or institution, and the likelihood of unauthorized or inadvertent disclosure. While in the U.S. there is no legal category for "sensitive data," regulations in other countries list information such as political opinions, religious beliefs, and sexual orientation under this category.[8]

Sensitive data does not inherently have to include identifying information to be classified as such. For example, genetic sequences alone are not considered PII under the HIPAA Privacy Rule but, because the genetic information may be placed in repositories and data banks and may contain an associated identifier for cataloging purposes, the information may be considered sensitive data in some research contexts.[9]

It's important to note that outside of human subjects research, the terms "sensitive", "confidential", and "private" may carry different or additional meanings than those described here. Sometimes these terms are also used interchangeably. For the most part, in human subjects research, data that are to be treated as "sensitive" or "confidential" are labeled as such because the research team has collected or has access to individually identifying information and, with the exception of participant consent, the research team has promised not to disclose the data in a way that re-identifies an individual. Data labeled as "private" in the HSR world typically requires that the research team seek permission from the participant to access their information for research purposes in the first place. Figure 1 showed the difference between data types in terms of the spectrum of identifiably of an individual. Figure 2 shows how the status of data can be understood in terms of its classification as private, confidential, or sensitive.

[8] See Solove's "What Is Sensitive Data? Different Definitions in Privacy Law" (2014) for a list of information that various countries regard as sensitive data.

[9] According to Garfinkel (2015), "there is no scientific consensus on the minimum size of a genetic sequence necessary for re-identification. There is also no consensus on an appropriate mechanism to make de-identified genetic information available to researchers without the need to execute a data use agreement that would prohibit re-identification" (p. 37).

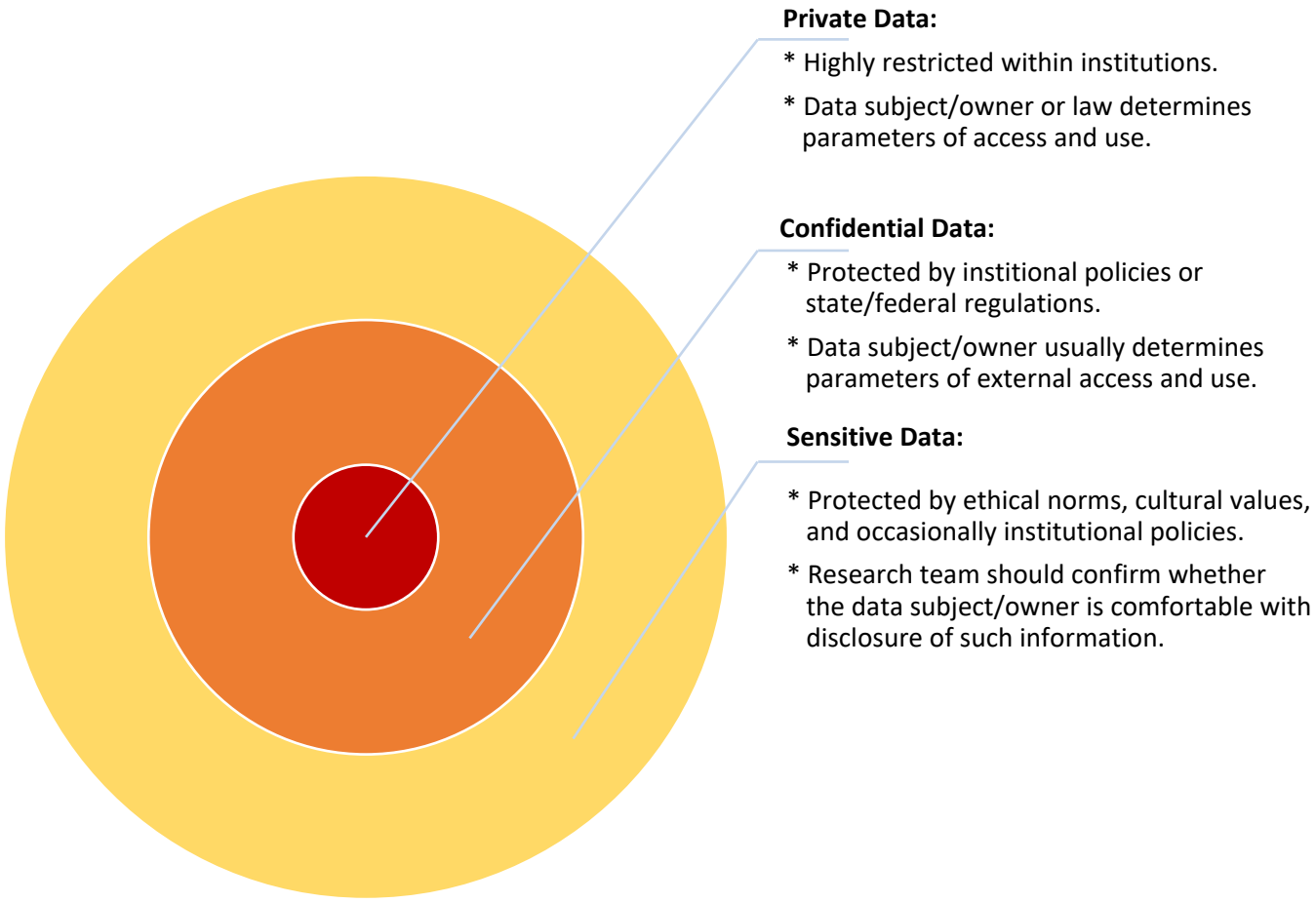

Figure 2. Relationship between private, confidential, and sensitive data in human subjects research.

Elements of an HSR Data Management Plan

1. Conduct a Data Inventory and Assess Level of Sensitivity

The first step in managing data effectively is to identify the type of data you expect to be working with and tailor your work to the challenges for the specific data set. Questions that should be addressed include:

- **What comprises HSR data for my research activity?**

This question should be addressed both before and after data collection. Prior to data collection, it is important for the PI to begin with making decisions about what data elements are needed to answer the research questions. PII should only be collected if it is relevant and necessary.

SJSU Policy: Requires that the PI be responsible for choosing an appropriate classification label (level 1, 2, or 3), based on the SJSU Information Classification and Handling Cheat Sheet (2019), to be used by all research personnel who create, compile, alter, or procure information collected from human subjects. "With the exception of general business correspondence and copyrighted software, all externally provided information that is not clearly in the public domain must receive a SJSU data classification system label. The SJSU worker who receives this information is responsible for assigning an appropriate classification on behalf of the external party. When assigning a SJSU classification label, this staff member must preserve copyright notices, author credits, guidelines for interpretation, and information about restricted dissemination" (p. 8). This requirement applies to the data across its life cycle. Information that does not have a label is, by default, classified as level 2, internal use only (Ibid, 7). To help assist with this required data inventory, the Office of Research has created an excel spreadsheet, the Data Management Plan Template, that allows PIs to list and describe the data elements to be collected, identify their classification, and note any restrictions on use.

- **Who owns the data? If the research data reside in personally-owned devices, what rights does my institution have to those devices?**

SJSU Policy: Makes no clear distinctions of ownership for raw research data; however, SJSU policies surrounding intellectual property address some ownership issues for the finished products of research. The following three SJSU policies address the topic of proprietary information:

S96-11 Fair Use of Copyrighted Materials; Intellectual Property (1996)
F98-3 Intellectual/Creative Property (1998)
S18-5 Research, Scholarship, and Creative Activity: Advisor-Student Relationship, Sponsored Projects, and Proprietary and Confidential Information in RSCA (2018)

Research that receives external funding or significant support by the University may also be subject to contractual agreements that require joint ownership and/or sharing of raw research data. In addition, investigators may develop collaborations where there is an expectation of joint ownership (whether formalized or not) by collaborators and co-PIs.

The concept of ownership of raw research data is worth diving into. Data privacy law scholar, Lothar Determann (2018), refers to the "landscape of interest" in formulating his argument that under current U.S. legal definitions data ownership cannot (and should not, in his opinion) be claimed by anyone; the nature of data, he points out, "implicates various legal positions, interests and options for parties interested in the data that are regulated in a considerate, nuanced and balanced fashion under laws outside the property law realm" (p. 4). Indeed, HSR is made possible by the contributions of multiple stakeholders, including the participants of the research who are the source of data, the research team members who analyze the data, and other scholars and sponsors who may wish to build on knowledge generated by the data. For this reason, investigators are urged to view their position as data stewards rather than as data owners but with the understanding that in some cases they may exercise privileges traditionally conferred upon owners (for example, the ability to take research data with them if they transfer to a different institution). The glossary definitions provided by the CSU Information Security Manual (2010) for data owner vs. data steward are offered here for PIs to contemplate how their roles might overlap:

Data Owner: "Person identified by law, contract, or policy with responsibility for granting access to and ensuring appropriate controls are in place to protect information assets. The duties include but are not limited to classifying, defining controls, authorizing access, monitoring compliance with CSU/campus security policies and standards, and identifying the level of acceptable risk for the information asset. A Data Owner is usually a member of management, in charge of a specific business unit, and is ultimately responsible for the protection and use of information within that unit."

Data Steward: "(Also known as 'Data Custodian') An individual who is responsible for the maintenance and protection of the data. The duties include but are not limited to performing regular backups of the data, implementing security mechanisms, periodically validating the integrity of the data, restoring data from backup media, and fulfilling the requirements specified in CSU/campus security policies and standards."

- **Does my research include information that can be used to distinguish or track an individual's identity (e.g., name, ssn, biometric information)? Does my research include information that could be used in conjunction with other data elements to reasonably infer the identity of a research participant (e.g., a combination of gender, race, date of birth, geographic indicators, or other descriptors)? What are the specific direct and indirect identifying information to be collected?**

CSU Policy: Requires that aggregates of data be "classified based upon the most secure classification level. That is, when data of mixed classification exist in the same file, document, report or memorandum, the classification of that file, document, report or memorandum must be of the highest applicable level of classification" (CSU Information Security Manual, Section 8065.S001, 2013, p. 2).

- **Are the data retained in digital or paper format or both?**

- **Does my research include app or web development that collects user/participant data?**

Researchers who wish to conduct user testing on apps to be developed for widespread use must follow CSU and SJSU policies, which include requirements for proper documentation of the app and the supporting test and production environment, conducting a risk assessment, identifying mechanisms by

which users' identities will be tracked as well as mechanisms for redacting user data in the production stage, and applying appropriate security protocols for protected data. PIs are instructed to refer to the following resources for complete information about app and web development requirements:

SJSU Cheat Sheet: Web Application Development (2015)
SJSU Standard: Application Service Provider Security Requirements (2015)

2. Determine Where Data Will Be Stored

Where data can be stored depends on the level of identifying information, if any, that is in possession of the research team and the security requirements that are needed. Questions that should be considered include:

- **Will the data reside in institutional/departmental data centers, on personally owned devices, in the cloud, or in third party databases?**

Storing data on personal devices exposes a PI to greater liability if the data are inadvertently exposed. Unless an external vendor has been properly vetted, protected data (e.g., level 1 or level 2 classification) should be stored on institutional devices or servers where access is limited by technical and procedural controls. Limiting data storage to campus devices and services also mitigates additional choices that researchers must make about storage when campus resources are not used, such as:

- **Is the data backed-up with a mechanism for recovery? Where are the back-ups stored?**

- **If using a free cloud storage service, is it truly free or does the vendor have access to data in a way that negatively affects the privacy of human subjects? Is the storage solution suitable for research data? What are the risks and the costs? How does a cloud storage service safeguard customer data? Does the cloud service provider store data outside the country? Does the cloud service provider have a reasonable mechanism for retrieving data once you no longer require their services or they go out of business?**

CSU Policy: "Employees must not store or transmit protected University data using services hosted by third parties which do not have a contract in place with the campus or its Auxiliaries, such as personal cloud accounts" (CSU Information Security Manual, Section 8065.S003, 2017, p. 1).

- **Will data be stored on a portable device (including a mobile phone) or removable media? How/where will such devices be stored?**

In general, protected data must not be stored on mobile devices unless effective security controls, such as password protection and encryption, have been implemented. Level 1 protected data may not be stored on mobile devices.

- **Where will the data be used?**

- **Where will the key for matching coded data back to the original unique identifiers be stored?**

- **Are duplicate copies of the data stored in different locations, and how will duplication be managed through the lifecycle of the data?**

3. Determine Who Will Have Access and Levels of Responsibility

People and entities with potential access to HSR data include: lead investigators, research team members (including student assistants), internal and external collaborators (including IT support, statisticians, and other consultants), funding agencies, supporting institutions, peer review researchers, editors, publishers, and data repositories. The following are some general tips for managing access to data. For more detailed information, PIs are advised to review SJSU's Access Control Standard (2017) to understand practices and policies related to access controls for level 1 and level 2 data.

- **Keep in mind the principles of separation of duties, need-to-know, and least privilege when considering who will have access to HSR data.**

For example, research personnel should only be given access to the data they need to carry out their assigned role on the research team; sensitive data that few personnel have access to should not be stored in the same location as other types of data used by other research personnel without additional protections in place for the data.

- **Decide and document the level of access that each of these groups will have, including access to: identifying or potentially identifying information, confidential or sensitive info, files, devices, databases, passwords, coding keys, and decryption keys.**

A spreadsheet, such as the Data Management Plan Template, provided by the Office of Research, can be used to track all stakeholder access.

- **Consider having key personnel sign an annual confidentiality pledge.**

- **Make sure to immediately revoke access privileges and update access tracking documentation when someone is no longer part of the research team.**

- **Assess that best practices and University policies for data security will be followed when outsourcing work to vendors.**

Educause advises that "due diligence should be conducted to determine the viability [of the vendor]. Consider such factors as reputation, transparency, references, financial (means and resources), and independent third-party assessments of safeguards and processes, particularly if you outsource the de-identification process" (Guidelines for Data De-identification or Anonymization, 2015).

- **Do not lend computers, phones, or any other equipment that contain sensitive HSR data, and report lost or stolen devices immediately.**

- **Do not provide access to SJSU systems to third parties without prior permission from IT.**

Unless otherwise stated under the data management plan in the IRB application, the PI is considered to be responsible for securing the research data and protecting its integrity and confidentiality. This also includes assigning various levels of access to all stakeholders. If there are multiple co-Is, all are equally responsible unless the data management plan indicates otherwise. Human subjects agree to participate in research with the expectation that the security and confidentiality of their data will be protected. PIs are responsible for ensuring this by vetting and educating research personnel and collaborators. This can

be done through security screening, training, and using confidentiality pledges (Institute of Educational Sciences, NCES, 2010, p. 9).

For collaborative research with outside institutions, the responsibility for proper data management should be noted in a collaborative agreement or a data use agreement (DUA). For example, "a DUA could prohibit a recipient of de-identified data from attempting to re-identify the data subjects, from linking to external data, or from sharing the data without permission" (Garfinkel, 2015, p. 5).

At SJSU, any institutional agreements must go through the proper channel for approval and signature. PIs are not allowed to sign on behalf of the institution. The SJSU Contracts Office and the SJSU Research Foundation are two entities that can assist with various contractual agreements.

4. Determine Level of Security and Level of De-Identification

"We live in a world where security by obscurity is no longer sufficient" (Nelson, 2015, p. 10).

Administrative, physical, and technical, safeguards are security controls that are intended to protect critical data assets, such as HSR data that contain identifying information, against loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure (Institute of Educational Sciences, NCES, 2010, p. 8). PIs should select a strategy for safeguarding HSR data that is appropriate to the level of risk of exposure. A combination of safeguards, as defined below, are generally required to protect PII.

Administrative safeguards. An institution's operating rules for the conditions of access and use to data; procedures for protecting PII when it is in the possession of authorized users; procedures for ensuring destruction of copies of records at the end of a period of authorized use; and systematic procedures for educating members of the institution about the operating rules (Ibid.).

Examples: Data use agreements; system-wide confidentiality agreements or pledges; non-disclosure agreements; information security and privacy training (including training provided by the Collaborative Institutional Training Initiative (CITI), to which SJSU subscribes, for research personnel conducting HSR).

Physical safeguards. Physical measures to protect data from unauthorized access.

Examples: Designation of security zones; entry controls; locked storage space, office, or building; securing devices or media to permanent furniture or structures within a space; keeping materials out of sight; use of screen filters; not using portable computers to store remote access credentials.

Technical safeguards. The technology and the procedures for its use that protect data and control access to it.

Examples: Password protection, multi-factor authentication, encryption, storage on non-network devices, de-identification.

CSU and SJSU Policy: All devices, including workstations, external drives, memory sticks, and mobile devices which store Level 1 protected data must be: 1) encrypted using campus approved encryption methods, and 2) tracked and managed via the campus inventory process (CSU Information Security Manual, Section 8050.S100, 2014, p. 1; SJSU Access Control Standard, 2017, p. 17; SJSU Asset Control Standard, 2015, p. 5-6).

De-identification is listed under technical safeguards since technical processes are often employed to transform the data. But, as guidance from the National Institute of Standards and Technology points out, “de-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset…The decision of how or if to de-identify data should thus be made in conjunction with decisions of how the de-identified data will be used, shared or released, since the risk of re-identification can be difficult to estimate” (Garfinkel, 2015, p. 1). Despite the difficulty of quantifying the risk of re-identification, one of the primary risks to consider is the cumulative threat of re-identification from all previous data releases and other reasonably available information, including publicly-available directory information and de-identified data releases from scholarly records. In particular, care should be taken to monitor new releases of de-identified individual-level data that are coded and stored in data repositories (U.S. Department of Education, PTAC, Data De-identification, 2012, p. 3).

It is the role of the IRB to examine the de-identification techniques and whether the research proposes an acceptable tradeoff between the potential for re-identification and the potential utility and beneficence of the data collected. Table 2 describes common de-identification techniques and the pros and cons of each.

Table 2 *Data De-Identification Techniques*

Name of Technique	Description / Examples	Pros	Cons
Redaction	**Erasing or expunging** sensitive data from a record.	Reduces risk if data are disclosed inadvertently or through unauthorized access; useful when the erased data elements are not needed for analysis (typical with direct identifiers).	Not effective if done improperly (e.g., if the erasure can be reversed or if enough indirect identifiers remain).
Suppression	**Removing** data (e.g., from a cell or row in a table, or data element(s) in a record) prior to dissemination to prevent the identification of individuals in small groups or those with unique characteristics. Examples: • Suppressing the value of a single field, such as a field in a patient record containing a very rare disease. • Not reporting observations for those patients where the number of patients for any combination of zip code, age, and diagnosis is below a given threshold (e.g., 5 people).	Useful when multiple indirect identifiers pose a risk for re-identification. More easily done with tabular data. Helpful when presenting analysis of findings to the institution that provided the data. Helpful in public health reporting.	May result in minimal data being produced for small populations, and it usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals). Can be difficult to perform properly. Is less likely to be effective if there are additional data available elsewhere.
Blurring: • Aggregation • Generalization • Pixelation	**Reducing precision** of data **by combining** one or more data elements. **Aggregation:** combining individual subject data with a sufficient number of other subjects to disguise the attributes of a single subject (e.g., reporting a group average instead of an individual value).	Minimizes risk of identification by focusing on collective data rather than individual data. Useful for "big picture" analyses.	Decreases reliability of data and increases potential for false conclusions. Aggregation: may not be possible with a small pool of subjects.

Name of Technique	Description / Examples	Pros	Cons
	Generalization: collecting or reporting values in a given range (e.g., using age or age-range instead of date of birth); including individual data as a member of a set (e.g., creating categories that incorporate unique cases); or reporting rounded values instead of exact amounts. **Pixelation:** modifying or obscuring visual information (e.g., blurring out faces in a photograph).		Generalization: unhelpful for case studies or in situations where details or specificity enhance findings. Pixelation: technology exists to reverse such modifications; other factors in a photo can lead to re-identification, such as setting and clothing.
Masking: • Pseudonymization Coding • Perturbation Randomization • Swapping Shuffling • Scrambling Encryption • Noise Differential Privacy	**Replacing** one data element with either a random or made-up value, or with another value in the data set; can be done manually or by using an algorithm. **Pseudonymization/Coding:** replacing a real name with a made up name or a real value with a made-up value. **Perturbation:** replacing sensitive info with realistic but inauthentic data or modifying original data based on predetermined masking rules (which may include randomization). Example: an algorithm which replaces the date of birth of subjects. **Swapping/Shuffling:** data for one or more variables are switched with another record, so that the data user does not know whether the real data values correspond to certain records (i.e.,	Attempts to retain the functional usability of the data while concealing information that could lead to identification. Pseudonymization/Coding: allows for a unique descriptor to trace data across multiple records; useful for multiple data instruments. Perturbation: reduces the likelihood of reverse identification. Swapping/Shuffling: useful for creating data sets for software testing where fields must be present and have realistic looking values.	Can decrease accuracy of computations in some cases, affecting validity of data. Techniques may be ineffective for small data sets. Algorithms used for masking can be reverse engineered.

Name of Technique	Description / Examples	Pros	Cons
	all the values in the data set are real, but are assigned to the wrong people). **Scrambling/Encryption:** data are algorithmically scrambled and only those with access to the appropriate key can view the encrypted data. **Noise/Differential Privacy:** statistical technique that introduces errors by randomly misclassifying values of categorical variable(s).	Noise/Differential Privacy: allows for quantification of potential privacy loss, enabling a more accurate risk assessment; useful for large data sets.	
Subsampling	**Releasing either a representative or random subsample** of data instead of an entire data set.	Minimizes risk of identification by reducing the amount of data reported.	May not yield representative and generalizable estimates of a study's overall subject population.

Table based on text combined from U.S. Department of Education, PTAC (2012) and Nelson (2015).

5. Determine How Data Will Be Shared and Disseminated

While item #3 in the data management plan involves determining the level of access to data for research personnel, here the PI must consider the best approach for sharing the research findings with the public (and with data repositories available to the public or to scholars) so that the research participants' identities are adequately protected. Item #s 4 and 5 in the data management plan are related since the level of de-identification will determine what data can be shared and vice versa. Some additional challenges that PIs should consider include:

- **The potential for re-identification when research involves a small number of subjects known to the researcher and/or the community to which the subject belongs.**

The more contextualized the information that is reported with indirect identifiers, the greater the risk of re-identification of the research participants. This is a risk that may need to be acknowledged during the consent process. The National Center for Education Statistics provides a relatable example that can occur in educational research: "a combination of data elements within student education records might reveal that there is only one student in a specific grade within a school with a set of observable characteristics who experienced a negative academic outcome" (Institute of Educational Sciences, NCES, 2010, p. 3). Thus, it would be appropriate for a PI to choose to drop or aggregate the data point in this scenario to avoid stigmatization of the individual student.

An analysis of whether reporting indirect identifiers poses a risk to human subjects should consider the likelihood of identifying an individual based on data elements reported in the research findings as well as the likelihood of readers being able to link the research data elements with information in external databases, whether they are public records or confidential data systems to which potential readers have access (Ibid.).

- **The potential for re-identification by friends, family, or associates of the subject.**

Research that enrolls friends and family members to discuss a sensitive topic may result in discord between individuals if the research team does not take into consideration the dynamics among the group of participants. For example, a qualitative study that attempts to understand the challenges of caregiving for the elderly may seek to interview both caregivers and those being cared for. Participants in this scenario may learn unfavorable information about individuals with whom they have a close relationship if they read the research results. Researchers should mitigate this possibility first with their research design. A study could, for example, focus on reporting best practices for caregiving in the elderly population. As well, recruitment procedures may need to include consensus and trust building between the research team and the group of participants, in addition to thorough disclosure during the individual consent process about who will be approached to participate and what they will be asked.

- **The potential quantity of information in the public domain about a specific individual.**

The more info there is in the public domain about an individual, the easier it will be to re-identify them as a research subject (e.g., if the subject is a community leader, a renowned professional, a celebrity). Depending on the nature of the research, the only feasible option may be to disclose the

identity of the individual -- with their consent -- because complete confidentiality could not be guaranteed anyway.

- **Open records access provided by some data repositories may conflict with the data management strategies put in place by the research team.**

PIs must understand the terms and conditions of use that repositories apply to data in their possession and must not share the data if the terms of use conflict with what was conveyed to the subjects and what was stated in the data management plan of the IRB protocol. As when contracting vendors to conduct work on behalf of the research team, PIs should fully vet the policies and practices of the data repositories to which they entrust data. Likewise, PIs who conduct sponsored research should be aware of the dissemination requirements of the sponsor, particularly the requirements of state and federal funding agencies that favor the open sharing of research data. Such requirements must be taken into account before data are collected to ensure that data elements that would expose subjects to unacceptable levels of risk are not collected.

- **Machine readability makes data easier to find online.**

- **Polices against dissemination of proprietary information may prevent data sharing.**

Though not typically related to the protection of human subjects, consideration must nonetheless be given to the protection of proprietary information, such as copyrighted material, information related to inventions and patents, and other intellectual property. Refer to the SJSU policies for proprietary information outlined under item #1 of the data management plan elements.

- **It is impossible to quantify future risk.**

Researchers can, at best, make educated guesses about the potential types of risks and their relative magnitudes when releasing data collected from human subjects. The very definition of risk is based on probabilities. Risk refers to the combined likelihood that a specific harm will occur and the probable magnitude of that harm.[10] When in doubt, researchers are advised to release only the minimum amount of information necessary to accomplish the research objectives identified at the outset. However, information security professionals may have something helpful to offer researchers in their risk analysis efforts: the practice of risk assessment mapping, a method that attempts to describe the consequences of a specific harm in codified language. Referring to such a map may help the research team judge whether or not a specific risk can be accepted. Table 3. offers a risk assessment map modified from the CSU Information Security Manual (2015). This sample map focuses on the risk of re-identification of a human subject based on a data set compiled and disseminated by the research team. Included with each severity rating – critical, high, moderate, low - is a corresponding action that should be taken; PIs can gauge the best course of action by first mapping the intersection between rows and columns that most resembles the risk scenario for their specific study and then referring to the actions listed in the column headings

[10] The federal Policy for the Protection of Human Research Subjects, for example, defines minimal risk as "the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests" (45 C.F.R. § 46.102(j), 2018).

Table 3 *Example of a Risk Assessment Map for Re-identification of a Research Participant*

Likelihood	Level of Severity: **Critical:** Severe impact. Immediate action needed. Example: loss of confidentiality with severe repercussion for subjects (e.g., damage to employability, financial standing, educational advancement, reputation; civil or criminal liability).	**High:** Significant impact. Seek support and resolve as quickly as possible. Example: loss of confidentiality with significant repercussion for subjects (e.g., embarrassment).	**Moderate:** Some impact. Take appropriate steps to report and mitigate the incident. Example: loss of non-sensitive data due to lost storage device.	**Low** Little impact. Generally ok to proceed with caution. Example: subject provides more information than the research team requested, which allows the research team to inadvertently be able to identify the subject.
Certain: - Likely to occur often. - Re-identification can easily be achieved by any member of the public.	**Critical**	**Critical**	**High**	**Moderate**
Probable: - May occur several times. - Re-identification can be achieved by highly motivated and sufficiently capable individuals with minimal effort (includes people who analyze or mine data with or without nefarious intent).	**Critical**	**Critical**	**High**	**Low**
Possible: - Likely to occur at some point. - Re-identification can be achieved by a known and limited number of people with detailed information (includes people who may know the subject well and are able to reasonably guess their identity).	**High**	**High**	**Moderate**	**Low**

Level of Severity

Likelihood

	Critical: Severe impact. Immediate action needed. Example: loss of confidentiality with severe repercussion for subjects (e.g., damage to employability, financial standing, educational advancement, reputation; civil or criminal liability).	**High:** Significant impact. Seek support and resolve as quickly as possible. Example: loss of confidentiality with significant repercussion for subjects (e.g., embarrassment).	**Moderate:** Some impact. Take appropriate steps to report and mitigate the incident. Example: loss of non-sensitive data due to lost storage device.	**Low** Little impact. Generally ok to proceed with caution. Example: subject provides more information than the research team requested, which allows the research team to inadvertently be able to identify the subject.
Seldom: - Unlikely but possible to occur. - Re-identification can be achieved only with a significant amount of guesswork, internal information, or expert technical knowledge. - Adequate controls are in place to protect the data.	Moderate	Moderate	Low	Low
Improbable: - So unlikely, it can be assumed occurrence may not be experienced. - Strong controls are in place to protect the data.	Low	Low	Low	Low

6. Determine How Data Will Be Transmitted/Transferred

Data may be transmitted digitally via wired, wireless, and cellular networks. As well, data may be physically transmitted via courier services. Questions that should be addressed include:

- **What methods, if any, will be used to protect data in transit?**
- **How is remote access to research data by external collaborators managed?**
- **Do the metadata in transmissions include information that needs to be protected?**

Consider the sensitivity level of the data to be sent via email, particularly over a wireless network which may be more vulnerable. Emailing unprotected PII or sensitive data may pose a high-security risk. The Privacy Technical Assistance Center at the U.S. Department of Education recommends alternative practices to protect data in transit, including "mailing paper copies via secure carrier, de-sensitizing data before transmission, and applying technical solutions for transferring files electronically (e.g., encrypting data files and/or encrypting email transmissions themselves; strong password protection and not storing and transmitting passwords in clear text)" (PTAC, Data Security Checklist, 2015, p. 3).

7. Understand Obligations for Protecting Data When Traveling

Researchers may forget that there are additional challenges to protecting data when traveling. Traveling abroad may complicate best efforts to protect data, as devices storing the data may be subject to international and domestic customs inspections, as well as surveillance threats from private entities. In general, sensitive data that contain PII should not be taken during travel intended for personal/non-professional reasons. SJSU policy for Research, Scholarship, and Creative Activity states that "RSCA team members may travel with confidential information to a location on campus or outside the campus, but team members must receive permission to do so from the PI" (SJSU, Policy S18-5, 2018, p. 8). While domestic and international regulations governing whether a person may export, import, or transfer data to and from specific institutions and countries is beyond the scope of this document, some general questions are provided below.

- **Does my research involve controlled information subject to U.S. export/import control regulations? If so, what are my responsibilities? How do export/import controls affect my research activity? How do I comply with the regulations?**

U.S. export/import control laws – principally concerned with the of disclosure of information related to any controlled technology or the material transfer of such technology to foreign countries or foreign nationals – do not affect the typical SBER research endeavor. Researchers whose data collection from human subjects or data sharing between research collaborators may include information about critical technologies; technical data/software code; equipment; chemicals/biological materials; and other materials, information, and services that relate to national security are advised to seek further guidance from SJSU's Office of Research.

Even when export/import regulations do not apply, it is prudent to consider the following questions when traveling:

- **Should I access emails or other communications that are housed on a U.S. server while out of the country?**

- **What data should I be concerned about taking with me (e.g., trade secrets, proprietary information)?**

- **What devices do I need to be concerned about (e.g., laptop, smartphone, mobile storage devices/thumb drives)? Should I use my laptop or get an institutional loaner?**

- **Can I decline to provide the password if asked for it by a customs or government officials? Does full disk encryption provide enough protection?**

Consult SJSU's Office of Research page on International Travel Guidance for SJSU-specific requirements and information.

The U.S. Department of State – Bureau of Consular Affairs provides information about specific countries and general guidelines on its international travel page.

The U.S. Department of Health and Human Services - Office of Human Research Protections provides an international compilation of human research standards, though researchers are advised to check local resources for potential variations and updates.

8. Develop a Retention Plan

When discussing record retention, it is important to note that we refer to applicable retention policies for HSR records that contain PII. Retention plans are not applicable to research data that have been stripped of all identifying information. Best practice dictates that research data, stripped of PII, be held for as long as necessary to ensure proper auditing of the research activities, when needed, and to ensure that future researchers are able to replicate the study. Investigators must be mindful of the distinction between records containing PII and research data in general when filling out the data management plan in the IRB application. IRB members have noted the erroneous but unfortunate frequent reference to "destroy all data once the research is complete" in many an IRB application.

Research records that contain PII should be maintained for as long as they serve a research function and, on occasion, for a specified period once the research is completed. Several policies and regulations apply to HSR records as follows:

CSU retention policy for research records is a *minimum* of 3 years (CSU Records Retention and Disposition Schedules, Section 10.6 Research Data, 2018), but it can be more if: 1) the funder/sponsor requires it, or 2) deferment of disposal is needed for investigations (both internal and external), the legal discovery process, and reasonably anticipated litigation (SJSU Record Retention and Disposal Standard, 2016, p. 6).

The exception to the 3-year minimum CSU requirement only occurs if the IRB approves the destruction of PII sooner than the 3-year minimum. For example, PII that is only collected for the purposes of tracking participants should be destroyed immediately once the research goals have been

met. Likewise, sensitive PII that exposes participants to risk so long as it is maintained should be destroyed immediately upon completion of the research.

SJSU's Record Retention and Disposal Standard (2016) supports the notion that the IRB has the ability to override CSU retention schedules by stating "external requirements under state and federal laws or regulation and University grants or contracts override university retention periods, where applicable" (p. 6). In addition, CSU Executive Order 1031 (2008), having to do with implementation of record retention and disposition schedules, refers to the obligation of ensuring "compliance with legal and regulatory requirements while implementing appropriate operational best practices" (Section I. Purpose). Given the IRB's federally mandated role of minimizing risk to research participants, we view the IRB as having a role to play in determining the retention period for HSR records that contain PII.

The above cited CSU Executive Order 1031 also points to the required documentation for campus custodians subject to the CSU policy. This documentation standard can be applied as a rule of thumb by data stewards (including researchers): "for each record/information listed, the schedule will include a unique number/identifier, title, custodian, value, retention authority, and retention period" (Section II. Definitions, Schedule). The Data Management Plan Template, provided by the Office of Research, can be used to label individual data elements that contain PII according to the documentation requirements of the CSU.

9. Understand Proper Methods for Disposing of PII

Once a retention plan has been identified, the method used to dispose of PII (also known as sanitization) should be selected based on the level of sensitivity of the data and the level of risk that would be associated with disclosure of PII that are not destroyed properly. The data inventory step in your data management plan should be referred to at this stage to determine which data are subject to disposal.

Specific methods of disposal will depend on the media type and hardware on which PII are stored.[11] Disposal may also include transferring data to another entity which becomes responsible for it, such as an archive or records repository. Standards and best practices are summarized as follows[12]:

- **Create a formal, documented process for data destruction – proof that data are properly and consistently disposed – and require that any partners involved in the research follow it.**

- **Draft written agreements with third parties about what, when, and how PII will be disposed of when it is no longer needed.**

- **Employ commonly used physical methods such as cross-cut shredding, pulverizing, or burning to destroy paper documents, CDs, DVDs, and any magneto-optical disks.**

[11] See Appendix A of the National Institute of Standards and Technology: Guidelines for Media Sanitization (2014) for more detailed information on minimum sanitization recommendations for different storage media (the table listing the storage media begins on p. 27 of Appendix A).

[12] This list is a modified version of text taken from the U.S. Department of Education's Privacy Technical Assistance Center (PTAC) document, Best Practices for Data Destruction (2019, p. 7).

- **Use appropriate deletion methods for sensitive electronic data to ensure they cannot be recovered**.

Note that simple deletion and disk formatting are not completely effective or appropriate methods when disposing of sensitive data. Often, when a data file is deleted, only the reference to that file is removed from the media. The actual file data remain on the disk and are available for recovery.

SJSU's Electronic Data Disposition Standard (2015) states that "data stored on devices must be: a. Rendered unreadable before leaving the possession of the university. b. Rendered unreadable before transfer to another organization, either internal or external to the university, and prior to being reused or repaired. c. Kept in a location limited to authorized personnel while waiting to be processed to render the storage media unreadable" (p. 5). The standard also lists a couple of basic options for rendering electronic files unreadable:

1. Use a campus-approved physical media destruction service. According to the disposition standard posted on its website, IT Services will accept delivery of equipment and dispose of data via a physical media destruction service.

2. Use a software-based, Department of Defense (DOD) approved disk-wiping utility[13] such as: Darik's Boot and Nuke (DBAN), KillDisk , Apple Disk Utility, Mobile Device "factory wipe" feature, and any other software package approved by the DOD.

Contact SJSU IT's security office for more information about data disposition tools: security@sjsu.edu.

- **Address in a timely manner sanitization of storage media which might have failed and need to be replaced under warranty or service contract.**

Many data breaches result from storage media containing sensitive information being returned to the manufacturer for service or replacement.

- **Remember: data may be blocked from routine disposal procedures due to a litigation or legal hold, audit, or public records request.**

Consider applying for a Certificate of Confidentiality (COC) through the National Institutes of Health if you will be collecting sensitive HSR data that could potentially lead to the incrimination of participants or that could compromise their welfare or safety if the data were disclosed. COCs protect identifiable research information from compulsory disclosures, such as court orders and subpoenas.

10. Understand Potential Threats to Confidentiality and Privacy of Subjects

Unfortunately, most data loss is due to the negligence of the individuals using the data rather than hacking (Nelson, 2015). An incomplete data management plan, poor data de-identification techniques, and careless reporting of data elements or statistics can pose risks to both individuals and groups. These risks include loss of privacy, identity theft, financial loss, embarrassment, stigmatization, discrimination,

[13] This method does not apply to Solid State Disk Drives (SSDs); SSDs are a type of mass storage device similar to a hard disk drive, but they use flash memory instead of magnetic discs. Hence, the methods used for wiping magnetic disks do not work on SSDs. A typical example of a SSD is a USB flash drive/thumb drive. The best way to ensure data disposition on SSDs is to physically destroy them.

emotional distress, and unnecessary costs to personal time and finances. Guidance from the National Institute of Standards and Technology points out that group harms can be experienced by an entire class of individuals, including individuals whose data do not appear in the dataset, because of inferences that can sometimes be drawn from statistical data. "For example, if a specific demographic group is well represented in a data set, and if that group [e.g., patients] has a high rate of a stigmatizing diagnosis in the data set, then all individuals in that demographic may be stigmatized, even though it may not be statistically appropriate to do so" (Garfinkel, 2015, p. 13).

Apart from being mindful of how data are processed and reported, investigators should also be aware of potential threats to data in order to adequately prepare a plan to best mitigate likely threats and to develop an incident response plan if a breach occurs. General threats to data include:

- Tampering or theft of intellectual property or government-sponsored research
- Alteration, damage, or loss of sensitive research data
- Unauthorized access or use of sensitive research data
- Improper disposal of digital media containing sensitive research data
- Sharing passwords and/or system access codes
- Unauthorized release of sensitive research data or product information
- Re-identification by linking research data with other identifying datasets (including publicly available data sets)
- Inferential disclosure of sensitive information from statistical properties of the released data
- Pseudonym reversal, especially if pseudonyms or codes are derived from identifying information

11. Develop an Incident Response Plan

While the impacts of a data breach to research participants have been noted in the previous section, there are also adverse consequences to investigators and institutions, such as: increased legal liability; loss of revenue, grants, gifts, and donations; bad publicity and damage to reputation; loss of public trust; and increased regulation, sanctions and/or legislation. Given the seriousness of some of the consequences of a data breach, the time to plan is not after a data breach has occurred. At a minimum, investigators should be prepared to follow the steps below promptly if they experience a data breach:

1. Identify that a breach has occurred

Make sure that all research personnel are trained to report a security incident to the PI immediately; this includes inadvertent but unauthorized access to HSR data by individuals outside of the research team. There may be different definitions of "data breach," but many of them do not depend on any harm occurring in order for the definition to apply. For example, if a laptop containing personally identifying information of research participants is lost, the incident would still be considered a breach even if no one acquires and uses the information for their own purposes. Likewise, some state and federal regulations may include the number of records affected as part of the definition of data breach for which varying notification rules apply. The HIPAA Breach Notification Rule, for instance, requires different reporting requirements based on a 500 record threshold (HIPAA Journal, 2019).

2. Conduct an impact assessment

The initial report must include the nature of the potential breach and an estimate of the severity – i.e., the number or records and types of information at risk of exposure (CSU Information Security Manual, Section 8075.00, 2010).

3. Know when to report to the institution, legal counsel, or law enforcement

The SJSU Information Security Incident Management Standard (2015) states that "lost machines must be reported to the Information Security Office and the appropriate property forms completed. Stolen machines must be reported to University Police in addition to all procedures required for lost computers…Identification badges and physical access cards that have been lost or stolen--or are suspected of being lost or stolen--must be reported to University Police and Facilities Development & Operations immediately. Likewise, all computer or communication system access tokens (smart cards with dynamic passwords, telephone credit cards, etc.) that have been lost or stolen--or are suspected of being lost or stolen--must be reported to the Information Security Office immediately" (p. 8-9).

4. Know who to contact and the reporting structure

SJSU Policy for the Protection of Human Research Subjects (2017) states that PIs must report an adverse event, which includes a security breach or lost or stolen equipment containing HSR data, to the IRB via the Office of Research within no more than 7 calendar days. Use the Incident Report features in SJSU's cloud-based submission system for IRB proposals to make a report related to an IRB-approved study.[14]

Data breaches that include exposure of Level 1 or Level 2 protected data must also be reported to the campus Information Security Officer (ISO). Contact information for both the ISO and the IRB are provided in Appendix A of this document. The ISO and the IRB are required, in turn, to follow the CSU, state, and federal reporting policies and laws for data breaches and adverse events.

5. Know when and how to notify impacted research participants

PIs should not contact research participants until they have reported the incident to the IRB and have been provided instructions on how to proceed, unless it is to prevent imminent harm to participants. SJSU policy holds that public statements must be free of explicit details and should only be released via the chain of command: "SJSU staff must not publicly disclose information about the individuals, organizations, or specific systems that have been damaged by computer crimes and computer abuses. Likewise, the specific methods used to exploit certain system vulnerabilities must not be disclosed publicly" (SJSU Information Security Incident Management Standard, 2015, p. 7).

12. Be Aware of Any Legal and Contractual Obligations That Apply to HSR Data

Though not an exhaustive list, the links below have been compiled to point to laws and regulations that may affect data management in an HSR endeavor. Laws that are most likely to impact SJSU research are listed, with a summary of the information most pertinent to data management included.

[14] Adverse events are not limited to data breaches or loss of equipment containing HSR data. They also include research-related injuries, which encompass both physical and psychological harms. Researchers are also required to report unanticipated problems related or possibly related to the research that place participants or others at greater risk of harm than was previously known.

Federal laws that may affect your data management plan. FERPA and HIPAA, discussed below, are the main federal regulations that are likely to affect SJSU investigators. A more comprehensive list of federal regulations that have provisions for the protection of PII are provided in the National Institute of Standards and Technology: Guide to Protecting the Confidentiality of Personally Identifiable Information (2010) – see appendix G. The applicable laws listed in the NIST publication appendix G mostly affect federal agencies or non-academic institutions (e.g., financial instructions, retail industry). However, investigators are encouraged to peruse the list provided by NIST to determine when additional regulations apply, especially for research that includes collaboration with either a federal government agency, industry, or a foreign government or foreign nationals.

- **Family Educational Rights and Privacy Act (FERPA)** (1974): Applies to student educational records.

With the exception of directory information as identified by the SJSU Registrar's Office, student educational records are classified as level 2 data under the SJSU Information Classification and Handling Cheat Sheet (2019) and should be protected as noted by this resource.

FERPA follows a consent-based approach for access to individually identifying student educational records for research purposes. Investigators who are in a dual role as educational service providers are not automatically allowed to use student records for research purposes without consent of the student, even if they have access to the records as part of their privileges as an instructor or administrator.

- **Health Insurance Portability and Accountability Act (HIPAA)** (1996)**:** Applies to medical records collected and maintained by covered entities serving patients. A covered entity can be a health plan or insurance provider, a medical billing service provider, or a health care provider.

An individual investigator who is not employed by a covered entity and who seeks health information directly from research participants is not considered to be a covered entity. Likewise, SJSU as a whole is not a covered entity, while certain departments within SJSU, such as the Student Wellness Center and the Speech Therapy Clinic within the College of Education do meet the covered entity definition. Interestingly, the Student Wellness Center is subject to both FERPA and HIPAA. The latter law (HIPAA) applies to the health professionals providing services to students in the Wellness Center, while the former law (FERPA) applies to all other entities who wish to access student medical records, including researchers.[15] Hence, investigators may find themselves consulting FERPA consent requirements if they wish to access student medical records from the SJSU Wellness Center. To add a layer of confusion to the matrix of regulations, the SJSU Information Classification and Handling Cheat Sheet (2019) regards medical records as level 1, highly protected data, while regarding student records covered by FERPA as level 2 data. When in doubt, apply the highest consent and security standards for level 1 data to student medical records.

[15] Refer to the Joint Guidance (2008) provided by the Department of Health and Human Services and the Department of Education on HIPAA and FERPA for more details on how the two laws are harmonized in their applicability to student health records.

SJSU investigators are more likely to access medical records provided by external institutions that are covered entities. Unless there is a written agreement or contract between SJSU and the covered entity, investigators are required to get documented informed consent from patients in order to access medical records that contain PII. The HIPAA Privacy Rule (2000) lists 18 data elements that, on their own, are considered to be PII:

1. Names
2. Zip codes (except first three numbers)
3. All elements of dates (except year)
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate or license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

In order for data to be considered de-identified by HIPAA standards, the research team can select from one of two options: 1) the above 18 data elements must be removed from the medical record by the covered entity prior to sharing records with the research team, or 2) a person with appropriate knowledge of generally accepted statistical methods and scientific principles applies and documents a statistical analysis that shows the risk of re-identification is low.[16]

- **Title IX** (1972): Applies to employees at institutions receiving federal funding. SJSU employees are required to report sex- and gender-based discrimination or harassment – including sexual harassment and misconduct, sexual exploitation, dating and domestic violence, stalking, or retaliation to the SJSU Title IX and Gender Equity Office. This policy applies to SJSU researchers, who may receive disclosures from research subjects during the course of the research, regardless of whether the person has requested anonymity.

[16] For more guidance on the HIPAA Privacy Rule and data de-identification, see Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (U.S. Department of Health and Human Services, OHRP, 2015).

- **Children's Online Privacy Protection Act (COPPA)** (2000)**:** Applies to operators of commercial web sites and online services, requiring them to provide notice and get verifiable parental consent before collecting personal information from children under the age of 13.

This law applies to investigators who want to work with online vendors or businesses to obtain secondary data collected from children or to investigators who wish to create apps as part of the research endeavor that may collect data from children. At SJSU, investigators occasionally collaborate with those that provide educational apps and COPPA may apply to those collaborations.

- **Federal Rules of Civil Procedures – Title V Disclosures and Discovery** (2019): Applies to parties in federal civil litigation and requires disclosure of documents or gathering a deposition or testimony from individuals.

Investigators should be aware that research documents can be requested when their work or the work of their collaborators is entangled in a civil lawsuit. Such requests can interfere with the privacy protections promised to human subjects during the course of the research. If you or your collaborators plan on collecting sensitive HSR data that could potentially lead to the incrimination of participants or that could compromise their welfare or safety if the data were disclosed, consider applying for a Certificate of Confidentiality (COC) through the National Institute of Health (the COC option is available for non-funded research as well). COCs protect identifiable research information from forced disclosure, such as court orders and subpoenas.

SJSU Policy: The Information Security Incident Management Standard (2015) states: "Employees are prohibited from providing any SJSU records, or any copies thereof, to third parties outside of SJSU or to government officials, whether in answer to a subpoena or otherwise, unless the prior permission of the Chief Legal Counsel or Public Records Act Officer has first been obtained. Likewise, employees are prohibited from testifying to facts coming to their knowledge while performing in their official SJSU capacities, unless the prior permission of the Chief Legal Counsel has first been obtained" (p. 13).

California laws that may affect your data management plan. Unlike the constitution of the federal government, California's state constitution specifically lists privacy as an inalienable right. The state has enacted numerous sector-specific privacy laws; those that may affect HSR data management are described below. We also describe state mandatory reporting laws and civil discovery laws to alert investigators to potential compulsory disclosures of information that may be gathered as part of the research activity. California laws that are harmonized with FERPA and HIPAA or that are not related to data management are not included here.

- **Privacy Laws:** The California Attorney General's Office provides this page with links to all of the state's privacy laws in one easy-to-access location. Laws that are especially relevant to SJSU research are summarized below.

Information Practices Act of 1977 – Conditions of Disclosure : Applies to investigators who wish to obtain secondary data that contain PII from state agencies, including public schools. The provisions of the act that apply to investigators are the same as those covered in this handbook. The act covers IRB review requirements, and investigators should be aware of the strict imposition on IRBs that approval to access PII should be based on demonstrated need by the research team. The act is also cited

in California Welfare and Institutions Code (1965), which requires IRB review for the release of any PII held by county or state agencies in relation to clients who receive public social services or welfare.

California Consumer Privacy Act of 2018: Applies to research data that contain personal information that have been collected from a consumer during the course of a consumer's interaction with a business. The act outlines that the use of such research data must be: 1) Compatible with the business purpose for which the personal information was collected, 2) Subsequently pseudonymized and de-identified, or de-identified and in the aggregate, such that the information cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer, 3) Made subject to technical safeguards that prohibit re-identification of the consumer to whom the information may pertain, 4) Subject to business processes that specifically prohibit re-identification of the information, 5) Made subject to business processes to prevent inadvertent release of de-identified information, 6) Protected from any re-identification attempts, 7) Used solely for research purposes that are compatible with the context in which the personal information was collected, 8) Not be used for any commercial purpose, and 9) Subjected by the business conducting the research to additional security controls limiting access to the research data to only those individuals in a business as are necessary to carry out the research purpose.

Business and Professions Code – Internet Privacy Requirements (2003): Applies to commercial websites or online services that collect PII, requiring operators to post a privacy policy and outlining the requirements for the privacy policy. Includes the following subsections:

Privacy Rights for California Minors in a Digital World (2013): Restricts marketing and advertising of certain products or services to minors by operators of websites, online services, or mobile applications.

Student Online Personal Information Protection Act (2014): Prohibits online operators catering to K-12 students from targeted advertising, amassing user profiles, selling student information, and restricts operators from certain disclosures of PII.

Early Learning Personal Information Protection Act (2016): Prohibits online operators catering to preschool and kindergarten-aged children from targeted advertising, amassing user profiles, selling student information, and restricts operators from certain disclosures of PII.

- **Education Code 51513 – Sensitive Surveys and Tests** (1976): Prohibits educators from surveying or testing K-12 students on certain sensitive topics without explicit written parental consent. Sensitive topics include questions about the pupil's personal beliefs or practices in sex, family life, morality, and religion, or any questions about the pupil's parents' or guardians' beliefs and practices in sex, family life, morality, and religion. This law also applies to investigators conducting surveys and is more restrictive than the federal Protection of Pupil Rights Amendment (PPRA) (2000), which only requires parental opt out. PPRA lists additional sensitive information that would trigger the parental consent requirement: political affiliations; mental and psychological problems potentially embarrassing to the student and his/her family; Illegal, anti-social, self-incriminating and demeaning behavior; critical appraisals of other individuals with whom respondents have close family relationships; legally recognized privileged or analogous relationships, such as those of lawyers, physicians, and ministers; or income (other than that required by law to determine eligibility for participation in a program or for receiving

financial assistance under such program).

- **Education Code - Privacy of Pupil Records** (2015): Places restrictions on the kind of information schools may collect from students' social media profiles and outlines the requirements that must be fulfilled for schools that wish to collect this information.

- **Health and Safety Code – Birth, Death, and Marriage Records** (1995): Imposes restrictions on the release and use of certain PII in birth, death, and marriage records. Specifically requires IRB approval for release of state death records containing PII to researchers (note, this is more restrictive than the federal regulations for protecting human research subjects, which only apply to living individuals).

- **Child Abuse and Neglect Reporting Act** (1987) **and Elder Abuse and Dependent Adult Civil Protection Act** (1994)**:** California mandatory reporting laws that may apply to some CSU employees and CSU student investigators.

The CSU, in its Executive Order 1083 (2017), provides guidance for employees on California mandated reporting requirements for child abuse and neglect, including categories of mandated reporters, forms, and instructions.

- **Civil Discovery Act - Sections 2016 - 2036** (2004): Applies to parties in California State civil litigation and requires disclosure of documents or gathering a deposition or testimony from individuals.

As on the federal level, California has its own laws regarding discovery in litigation. Investigators are advised to consult legal counsel to understand requirements of federal or state discovery laws, and to follow SJSU institutional policy, as noted under Federal Rules of Civil Procedures discussed earlier in this section.

How funding agencies may affect your data management plan. If you will be applying for funding from a federal sponsor, they may have their own rules and regulations for research data. The SJSU Research Foundation processes grant submissions primarily for public and corporate funds and helps investigators ensure that a specific sponsor's requirements are met.

Most sponsors will not require anything more comprehensive in their data management plan than what is covered in this handbook; however, investigators should check with the sponsor to be certain. In addition, the SJSU Library offers a Research Guide on Data Management (among other excellent research guides) as well as access to DMPTool, an online tool for constructing a data management plan based on templates from multiple funding agencies.

Other requirements that may affect your data management plan.

- Requirements of private funders

- International research:

General Data Protection Regulation (2018): A comprehensive, multi-sector regulation of data belonging to EU citizens and residents. The law applies to any non-EU organization or institution that

does business in the EU. Many U.S. companies have adopted the requirements of GDPR.

International Compilation of Human Research Standards: Provided by the U.S. Office of Human Research Protections as a starting point for investigators who wish to conduct research outside of the U.S.

References

Biometrics. (n.d.). In *Wikipedia*. Retrieved March 18, 2019, from https://en.wikipedia.org/wiki/Biometrics

California Consumer Privacy Act of 2018, Cal. CIV. § 1798.100 – 1798.199 (2018).

California Department of Justice, Attorney General. (n.d.). *Privacy laws*. Retrieved from https://oag.ca.gov/privacy/privacy-laws

California Education Code. Privacy of Pupil Records, Cal. EDC. § 49073 – 49079.7 (2015).

California Education Code. Sensitive Surveys and Tests, Cal. EDC. § 51513 (1976).

California Health and Safety Code. Birth, Death, and Marriage Records, Cal. HSC. § 102175 – 102249 (1995).

California State University. (2008, February 27). *CSU executive order 1031*. Retrieved from https://www.calstate.edu/EO/EO-1031.html

California State University. (2017, July 21). *CSU executive order 1083*. Retrieved from http://www.calstate.edu/EO/EO-1083-rev-7-21-17.html

California State University. (2010). *Information security manual* (ICSUSM Section 8000). Retrieved from https://calstate.edu/icsuam/documents/Section8000.pdf

California State University. (2018, August 23). *Records retention & disposition schedules* (Section 10. Research & sponsored programs). Retrieved from: http://www.calstate.edu/recordsretention/documents/RSP.pdf

California Welfare and Institutions Code. Records, Cal. WIC. § 10850 – 10853 (1965).

Child Abuse and Neglect Reporting Act, Cal. PEN. § 11164 – 11174.3 (1987).

Children's Online Privacy Protection Act, 16 C.F.R. 312 § (2000).

Civil Discovery Act, Cal. CCP. § 2016 - 2036 (2004).

Collaborative Institutional Training Initiative (CITI) Program. (n.d.). *Research ethics and compliance training*. Retrieved from https://about.citiprogram.org/en/homepage/

Determann, L. (2018). No one owns data. *UC Hastings Research Paper No. 265*. Retrieved from SSRN: https://ssrn.com/abstract=3123957

Duncan, G.T., & Elliot, M., & Salazar-Gonzalez, J.-J. (2011). *Statistical confidentiality: Principles and practice.* New York: Springer-Verlag.

Early Learning Personal Information Protection Act, Cal. BPC. § 22586 – 22587 (2016).

Education Amendments Act of 1972, 20 U.S.C. §§1681 - 1688 (2018). Title IX (Education Amendments Act of 1972, 2018)

Educause. (2015). *Guidelines for data de-identification or anonymization*. Retrieved from https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/resources/information-security-guide/toolkits/guidelines-for-data-deidentification-or-anonymization

Educause. (2017). *Top information security concerns for researchers*. Retrieved from https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/resources/information-security-guide/toolkits/top-information-security-concerns-for-researchers

Elder Abuse and Dependent Adult Civil Protection Act. Mandatory and Nonmandatory Reports of Abuse, Cal. WIC. §15630 – 15632 (1994).

Emam, K.E. (2010). Risk-based de-identification of health data. *IEEE Security and Privacy*, *8*, 64-67. doi: 10.1109/MSP.2010.103

European Union. *General Data Protection Regulation (GDPR)*. (2018). Retrieved from https://gdpr.eu/what-is-gdpr/

Family Educational Rights and Privacy Act, 34 C.F.R. § 99 (1974).

Future of Privacy Forum. (2017, April 25). *A visual guide to practical data de-identification*. Retrieved from https://fpf.org/wp-content/uploads/2017/06/FPF_Visual-Guide-to-Practical-Data-DeID.pdf

Garfinkel, S. L. (2015). *De-identification of personal information* (NISTIR 8053). Retrieved from National Institute of Standards and Technology website http://dx.doi.org/10.6028/NIST.IR.8053

Health Insurance Portability and Accountability Act. Privacy Rule, 45 C.F.R. §160 and 164 (subparts A and E) (2000).

HIPAA Journal (2019, February 14). *Deadline for reporting small healthcare data breaches*. Retrieved from https://www.hipaajournal.com/march-1-2019-deadline-for-reporting-small-healthcare-data-breaches/

Information Practices Act of 1977. Conditions of Disclosure, Cal. CIV. § 1798.24 - 1798.24b (1977).

Institute of Education Sciences, National Center for Education Statistics, Statewide Longitudinal Data Systems Grant Program. (2010, November). *Data stewardship: Managing personally identifiable information in electronic student education records* (NCES 2011-602, SLDS Technical Brief 2). Retrieved from https://nces.ed.gov/pubs2011/2011602.pdf

Internet Privacy Requirements, Cal. BPC. § 22575 – 22579 (2003).

John Hopkins Medicine, Office of Human Subjects Research – Institutional Review Board. (2015). *Definition of limited data set*. Retrieved from https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/limited_data_set.html

Kissel, R., & Regenscheid, A., & Scholl, M., & Stine, K. (2014). *Guidelines for media sanitization* (NIST Special Publication 800-88, Rev. 1). Retrieved from National Institute of Standards and Technology website https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-88r1.pdf

McCallister, E., & Grance, T., & Scarfone, K. (2010). *Guide to protecting the confidentiality of personally identifiable information* (NIST Special Publication 800-122). Retrieved from National Institute of Standards and Technology website https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

Nelson, G. S. (2015). *Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification* (ThotWave Technologies, Chapel Hill, NC, Paper 1884-2015). Retrieved from the SAS Institute support website https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701 – 1777; *University of Colorado Law Legal Studies Research Paper No. 9-12*. Retrieved from SSRN: https://ssrn.com/abstract=1450006

Policy for the Protection of Human Research Subjects, 45 C.F.R. § 46. (2018).

Privacy Rights for California Minors in a Digital World, Cal. BPC. § 22580-22583 (2013).

Protected Health Information. (n.d.) In *Wikipedia*. Retrieved January 12, 2019, from https://en.wikipedia.org/wiki/Protected_health_information

Protection of Pupils Rights Amendment, 20 U.S.C. § 1232h (2000 & Supp. IV 2004).

San José State University. (1996). *Fair use of copyrighted materials; intellectual property* (Policy S96-11). Retrieved from http://www.sjsu.edu/senate/docs/S96-11.pdf

San José State University. (1998). *Intellectual/Creative property* (Policy F98-3). Retrieved from http://www.sjsu.edu/senate/docs/F98-3.pdf

San José State University. (2017). *Protection of human research subjects* (Policy F17-1). Retrieved from http://www.sjsu.edu/senate/docs/F17-1.pdf

San José State University. (2018). *Research, scholarship, and creative activity: Advisor-Student relationship, sponsored projects, and proprietary and confidential information in RSCA* (Policy S18-5) Retrieved from http://www.sjsu.edu/senate/docs/S18-5.pdf

San José State University, Finance and Business Services. (n.d.). *Contracts*. Retrieved from https://www.sjsu.edu/fabs/services/contracts/index.php

San José State University, IT Information Security Office. (2015). *Cheat sheet: Web application development*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Cheat_Sheet_Web%20_Application_Development.pdf

San José State University, IT Information Security Office. (2015*). Standard: Application service provider security requirements*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Application_Service_Provider.pdf

San José State University, IT Information Security Office. (2015). *Standard: Asset control*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Asset_Control.pdf

San José State University, IT Information Security Office. (2015). *Standard: Electronic data disposition*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/DataDispositionStandard.pdf

San José State University, IT Information Security Office. (2015). *Standard: Information security incident management*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Information_Security_Incident_Management.pdf

San José State University, IT Information Security Office. (2015). *Standard: Risk assessment program*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Risk_Assessment_Program.pdf

San José State University, IT Information Security Office. (2016). Standard: *Record retention and disposal*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/RecordRetentionDisposalStandard.pdf

San José State University, IT Information Security Office. (2017). *Standard: Access control*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Access_Control.pdf

San José State University, IT Information Security Office. (2019). *Cheat sheet: Information classification and handling*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Cheat_Sheet_Information_Classification.pdf

San José State University, IT Information Security Office. (2019). *Standard: Information classification and handling*. Retrieved from http://www.sjsu.edu/it/docs/security/policies-standards/Standard_Information_Classification_Handling.pdf

San José State University, Library. (2019). *Data management, resources in documenting, storing and preserving research data*. Retrieved from https://libguides.sjsu.edu/datamanagement

San José State University, Office of Research. (n.d.). *Export Control*. Retrieved from https://www.sjsu.edu/research/research-compliance/export-control/index.php

San José State University, Office of Research. (n.d.). *International Travel Guidance*. Retrieved from https://www.sjsu.edu/research/research-compliance/international-travel-guidance/index.php

San José State University, Office of Research. (n.d.). *IRB: Researcher Training*. Retrieved from https://www.sjsu.edu/research/research-compliance/irb/irb-researcher-training.php

San José State University, Office of Research (2019). *Data management template*. [Excel form] Retrieved from http://www.sjsu.edu/research/docs/irb-data-management-plan-template.xlsx

San José State University, Office of the Registrar. (n.d.). *Family Educational Rights and Privacy Act (FERPA)*. Retrieved from https://www.sjsu.edu/registrar/academic-records/ferpa.php

San José State University, Research Foundation. (n.d.). *Industry Agreements*. Retrieved from https://www.sjsu.edu/researchfoundation/resources/industryagreements.php

San José State University, Title IX and Gender Equity Office. (n.d.) https://www.sjsu.edu/titleix/index.php

Shostak, J. (2006, May). *De-Identification of Clinical Trials Data Demystified*. Paper presented at the PharmaSUG Conference, Denver, CO. Abstract retrieved from https://www.lexjansen.com/pharmasug/2006/PublicHealthResearch/PR02.pdf

Solove, D (2014, July 31). What is sensitive data? Different definitions in privacy law [Web log post]. Retrieved from https://teachprivacy.com/sensitive-data-different-definitions-privacy-law/

Student Online Personal Information Protection Act, Cal. BPC. § 22584-22585 (2014).

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Carnegie Mellon University, Data Privacy Working Paper 3). Retrieved from Data Privacy Lab website https://dataprivacylab.org/projects/identifiability/paper1.pdf

Tea Room Trade. (n.d.). In *Wikipedia*. Retrieved April 10, 2019, from https://en.wikipedia.org/wiki/Tearoom_Trade

Title V Disclosures and Discovery. Fed. R. Civ. P. 26-37. (2019).

U.S. Department of Education, Privacy Technical Assistance Center and the Student Privacy Policy Office. (2012). *Data de-identification: An overview of basic terms* (PTAC-GL). Retrieved from the Protecting Student Privacy Website https://studentprivacy.ed.gov/resources/data-de-identification-overview-basic-terms

U.S. Department of Education, Privacy Technical Assistance Center and the Student Privacy Policy Office. (2015). *Data security checklist* (PTAC-CL-2). Retrieved from the Protecting Student Privacy Website https://studentprivacy.ed.gov/resources/data-security-checklist

U.S. Department of Education, Privacy Technical Assistance Center and the Student Privacy Policy Office. (2019). *Best practices for data destruction* (PTAC-IB-5). Retrieved from the Protecting Student Privacy Website https://studentprivacy.ed.gov/resources/best-practices-data-destruction

U.S. Department of Health and Human Services, National Institutes of Health. (2019) Certificates of Confidentiality (COC) – human subjects. Retrieved from https://grants.nih.gov/policy/humansubjects/coc.htm

U.S. Department of Health and Human Services, Office of Human Research Protections. (2015). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Retrieved from

https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard

U.S. Department of Health and Human Services, Office of Human Research Protections. (2019). *International compilation of human research standards*. Retrieved from https://www.hhs.gov/ohrp/international/compilation-human-research-standards/index.html

U.S. Department of Health and Human Services and U.S. Department of Education. (2008). *Joint guidance on the application of the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to student health records*. Retrieved from https://studentprivacy.ed.gov/resources/joint-guidance-application-ferpa-and-hipaa-student-health-records

U.S. Department of State, Bureau of Consular Affairs. (n.d.). International travel. Retrieved from https://travel.state.gov/content/travel/en/international-travel.html