Guidelines for Data De-Identification or Anonymization

Last reviewed: July 2015

NOTE: For the purposes of this document, although there are subtle differences in their definitions, "de-identification" and "anonymization" will be considered synonymous terms. These terms refer to situations where personally identifying information is removed from data sets in order to protect a person's individual privacy. "Sanitization" is considered a media disposal term and refers to removing data from media storage devices. More information on sanitization can be found in the Guidelines for Information Media Sanitization.

Table of Contents

Purpose

This document outlines high-level definitions, key challenges and risks, recommendations, critical first steps, and resources for the implementation and use of de-identified or anonymized data. It does not contain specific technical methods for the de-identification of particular data sets. The document is written specifically with institutions of higher education in mind; however, these high-level issues are likely common to most organizations attempting data de-identification.

Definitions

Definitions with respect to concepts such as "de-identification," "anonymization," and "sanitization" are highly nuanced and context-dependent. Institutions of higher education are encouraged to define these concepts as they apply to local institutional policies, processes, and procedures in order to eliminate confusion regarding similar terms. The definitions below provide information about how terms may be used in different contexts.

These definitions are based in part on the IAPP's Glossary of Privacy Terms.

- *Anonymization*: The act of permanently and completely removing personal identifiers from data, such as converting personally identifiable information into aggregated data. Anonymized data is data that 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 underlying individual.
- *Data Handler*: Sometimes also called a "data processor." This is an individual who processes, handles, or otherwise uses data at an institution. With respect to de-identifying data, this is the individual who takes the original data and does the work to de-identify it.
- *Data Subject*: The term used to describe the individual who is the subject of a data record.
- *De-identified*: Without reference to health information, de-identification involves the removal of personally identifying information in order to protect personal privacy. In some definitions, de-identified data may not necessarily be anonymized data (as we have defined that term in this document). This may mean that the personally identifying information may be able to be re-associated with the data at a later time. In such cases, anonymized data is a particularized subset of de-identified data. In this document, "de-identified" and "anonymized" will be considered synonymous terms.
  This term is also understood as a health information concept as it relates to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Data is considered de-identified under the Privacy Rule when a number of specified data elements are removed. (45 C.F.R. §§ 164.502(d)(2), 164.514(a) and (b).) De-identified data is not regulated by HIPAA and may be shared without restriction. This concept is different from the HIPAA "limited data set" concept. A "limited data set," by contrast, is stripped of many categories of identifying information but retains information often needed for public health and research (such as birth dates, dates of treatment, and some geographic data). Entities covered by HIPAA may share a limited data set for research, public health and health care operations purposes permitted by the Privacy Rule, so long as all recipients are bound by a data use agreement with the originator of the data. (45 C.F.R. § 164.514(e).)
- *Sanitization*: Refers generally to the process of removing information from storage media such that data recovery is not possible. See Guidelines for Information Media Sanitization. "Data Sanitization" can also refer to the process of disguising sensitive information in information technology resources by overwriting it with realistic looking, but false, data. *For the purposes of this document, "sanitization" is considered a media disposal term.*

Top of page

Overview

The ability to collect and store information about individuals and their actions and habits is easier than ever before. Advances in information technology make the storage, cataloging, and use of such information trivial. Many educational institutions have stored both paper and electronic data about individuals, either through the direct collection of such data for organizational purposes or data stored as a result of the provision of services to individuals. Due to privacy concerns, oftentimes such data must be de-identified or anonymized before it is used or studied.

Educational institutions may have a number of reasons for using de-identified data for business, academic, or operational functions. For instance, data can be made available for institutional use, without identifying the underlying data subjects, for research purposes, institutional effectiveness studies, performance and operational studies, information technology security and operational reviews, and for public health purposes.

Other uses of de-identified data may require the ability to retain unique identifiers for individuals in the data set, without identifying the actual identity of the individuals. For example, a researcher may need to know that certain actions were all taken by the same individual, in order to form conclusions about how individuals use the data or service. A web site designer may want to determine how long individuals stay on the site, or how individuals traverse the site in order to find the information sought. Systems development, test, and training environments may require the use of data that simulates real production data, while not actually consisting of real data elements such as Social Security numbers. In such cases, de-identification processes are complicated by the need to replace unique identifiers such as Social Security numbers or IP numbers with alternate unique identifiers that cannot be used to identify the actual individual.

While de-identifying data is a useful step towards protecting privacy, the de-identified data can still carry a number of privacy risks. For instance, in some situations institutions may need to ensure that de-identified or anonymized data cannot be re-engineered to identify the underlying data subjects. This concern is not insignificant and there are a number of examples of purported de-identified industry data being released for research purposes that was subsequently found to be poorly de-identified or susceptible to re-identification of individual data subjects. For instance, in 2006, AOL released search log data on subscribers that had been intended for use with the company's newly launched research site. Although no personally identifiable data was included, privacy advocates found that individuals could be identified based on the detailed searches they conducted. Netflix also released movie ratings data that had been anonymized by removing personal details; yet, researchers were able to de-anonymize the data by comparing it against publicly available ratings on the Internet Movie Database. Thus, as part of a risk assessment when releasing de-identified data, consideration must be given to the likely possibility of the recipients of the data having access to other data sets or methods that would allow re-identifying the data. Clearly, releasing de-identified data sets to the public, as with the AOL and Netflix incidents, requires more careful consideration of this risk than releasing de-identified data to a researcher within your organization or other distinct entity, especially if an agreement can be documented as to what the researcher or entity can and cannot do with the de-identified data.

Institutions looking to address de-identification or anonymization of data are strongly advised not to proceed without partnering with their Institutional Review Boards. The IRB's primary function is to protect the privacy of individuals and the rigor of research protocols associated with human subjects. As such, they bring not only significant expertise to the issue, but a deep understanding of institutional processes as they affect research.

Top of page

Key Challenges and Risks

Before embarking on a data de-identification project, high-level challenges and risks must be identified to determine how to appropriately mitigate risks in the context of the proposed use of the data. The list of challenges included below is intended to help institutions identify their own unique issues regarding de-identified or anonymized data sets. This list is not intended to be complete, and not all challenges pose risks that would outweigh the benefits of the use of the data. It is important to review the challenges and risks against these benefits, and to identify strategies for reducing risks, before making decisions to de-identify and then release the data.

No regulation of de-identified data

While the United States has no one general law regarding the privacy of data, identified or de-identified, there are a number of different requirements and definitions for the data used in the various regulatory sectors. For instance, the federal Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Pub. L. No. 104-191, § 264 (1996), codified at 42 U.S.C. § 1320d), primarily protects the use of protected health information. The Gramm-Leach-Bliley Act (GLBA) (Pub. L. No. 106-102 (1999) protects some types of consumer financial data. In both of these laws, however, de-identified data is not truly regulated. In addition, de-identified, publicly available data does not constitute human subjects research as defined at 45 C.F.R. 46.102.

Institutions should consider how the provision of de-identified or anonymized data impacts their ability to comply with the reporting duties of various legal requirements. E-discovery, state data protection laws, and export control regulations may also need to be considered. (*Note: See the E-Discovery Toolkit fo r more information.*)

Lack of clear definition of de-identified or anonymous data

It is often not possible for an educational institution to declare for certain when a data set has or has not been de-identified. All organizations face significant challenges and risks in ensuring that their processes for de-identifying or anonymizing the personal identifiers in data sets are accurate.

The lack of a generally accepted overarching definition of what constitutes "personally identifiable information" ("PII") versus non- personally identifiable information ("non-PII") exists because it is not possible to reduce the issue to a simple listing of data elements. Information that enables an individual to be distinguished as a particular computer user is dependent on the context and the availability of other data sets that, when compared to the de-identified data set, could cause identification to occur regardless of whether the data is "personally identifiable" in the traditional sense.

The AOL and Netflix examples, in particular, show that there are situations in which the personal identifiers or other identifying information in de-identified or anonymized data can be recovered or reconstructed, when the data is released to those who have access to other data sets that might enable re-identification, and who have not agreed to appropriate terms of use of the data.

Paper-based vs. electronic data

The steps taken to de-identify data will differ based on the format of the data. De-identification concerns have arisen primarily because of the production of huge sets of electronic data, but data in paper format also may need to be de-identified. In such cases, methods such as using a black marker to obscure the identifiable parts of the document are not usually sufficient, as the information may still be legible. Physically cutting out the identifiable information is usually recommended. When paper documents are converted to images, the imaging software may allow for blacking out of data in a way that renders the area unreadable.

Types of de-identified or anonymous data

Educational institutions collect personally identifiable data as a product of doing business with students, faculty, staff, and outside parties. Educational institutions also collect data about their information technology systems and operations, in the form of logs, network traffic, web traffic, etc., which also may contain personally-identifiable data. There are different challenges and risks with de-identifying the data, depending on its type. Thus, it is important to first determine which types of data you will be working with, and to tailor your work to the special challenges for that type.

Special challenges with logs, network traffic, web traffic, etc.

The challenge in assuring that data is fully de-identified or anonymized is compounded when attempting to de-identify huge sets of systems operations data in unstructured formats. There are no search terms that can be reliably used to find and remove all potential instances of personally identifiable data (for example, names and addresses). Anonymizing tcpdump packet captures is extremely difficult to do because the packet contents reveal a great deal of information about the users. In flow dumps, even if address information is anonymized, traffic and pattern analysis would allow analysis that may be personally identifiable. In addition, there is currently a debate as to whether the IP Address, when it appears in log or traffic data, constitutes personally identifiable data. Some have chosen to truncate the last one or two octets of the IP address in order to avoid that debate; however, others believe this truncation is still not de-identified enough.

Need for re-identification and careful use of re-identification keys

In some cases, the de-identified data needs to eventually be re-identified, or, the de-identified data may need to retain the ability to track the activity of an anonymous individual in the data set. In such cases, de-identification processes are complicated by the need to replace unique identifiers such as Social Security numbers or IP numbers with alternate unique identifiers that cannot be used to identify the actual individual. The key for matching the alternate unique identifier back to the original unique identifier may need to be retained, yet highly secured from unauthorized access. Additionally, researchers often need to trend data - thus, anonymization keys need to be varied periodically or else it becomes easier to recover or resolve network structure - a possible security impact to the institution that is unrelated to personally identifiable data.

Balancing risk with value

The consequences of poor data de-identification or anonymization can be severe: individuals can be personally identified, perhaps with respect to sensitive or embarrassing financial, medical, or other habits or activities. If network or other operational data is poorly de-identified or anonymized, a person external to the institution may be able to map the institution's network infrastructure. However, de-identification and anonymization must be balanced with the value of the data. In the network example, totally anonymized data (such that an institution's network topology becomes featureless) minimizes the value of the data. It has been said that "Data can either be useful or perfectly anonymous but never both." If the value is high, organizations can often identify and implement strategies to reduce the risk to an acceptable level, so that the data can be utilized.

The status of the requestor also factors into the risk. If the requestor is a member of the organization providing the data, this is generally less of a risk than providing the data to an external party. Can the requestor be required to sign an agreement specifying how the de-identified data may or may not be used? This is a typical strategy used to reduce the risk. It is important to ensure the institution has a method for evaluating the value against the risk, and has strategies for reducing the risk to acceptable levels, so that data can be utilized for research, business, academic, operational, and other purposes when there is significant value.

For more information on risk analysis, see the Risk Management chapter.

Handling and use considerations

There may be a number of persons involved in a transaction involving de-identified and anonymized data. These persons may take on many roles: the data handler who takes the full data set and de-identifies it; the receiver of the de-identified data; the external researcher who manipulates de-identified or anonymized data sets or combines those data sets with outside information. Safeguards must be considered with respect to how such data sets will be used and by whom. Considerations include:

- Human resources safeguards: How do institutions ensure that both institutional and external data handlers minimize risks to the privacy, security, and integrity of the data? How does the provider manage additional individuals who may be given access to the data?
- Receiver trustworthiness: How does an institution establish an adequate level of trust in the receiver of the data? How do we ensure our trust boundaries do not extend farther than intended when giving the data to the receiver? Does receiver intended use of information conflict with institutional privacy principles?
- Co-mingling with other data: What are the potential for and the consequences of information re-identification resulting from co-mingling the de-identified or anonymized data set with other information the receiver has access to? An example might include Google Analytics, where analytics data such as IP address could be compared to other data that Google retains.
- Responsibility/liability for breaches: What is the relative liability for privacy breaches accepted by the receiver and retained by the university? How will liabilities related to privacy breaches be shared between the receiver and the university? What response plan will be followed if a privacy breach occurs? How is the data owner notified? What leverage do we have if the receiver acts inappropriately with the data?
    - Note that in some instances, institutions may even have notification responsibilities with respect to de-identified data, such as in the instance of protected health information under HIPAA. Under the HITECH Act, breach notification may be required if, based upon rational reasoning and analysis, it is determined that protected health information elements, individually or in combination, could point to a specific individual(s). (Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA) (Pub. L. 111-5).)
- Confidentiality/privacy: What are the privacy risks and/or open records consequences of de-identified or anonymized data sets and/or service involved? Must institutions provide an opt-out for individuals in an original, identified data set who do not wish to have their data included in de-identified data sets?

Data Classification

Institutions have many different regulations that they must follow. Data classification is often a security requirement under these regulations. For instance, HIPAA, the Family Educational Government Rights and Privacy Act (FERPA) (Pub. L. No. 93-380 (1974), codified at 20 U.S.C. § 1232g), and the Department of Health and Human Government Services (HHS) regulations regarding protection of human subjects (Title 45 CFR Part 46) all require classification of data. Classifying data appropriately is an important step in properly protecting it. Understanding the classification requirements for a set of data can help an organization determine whether data should be de-identified and/or anonymized when it is used for certain purposes.

More information on classification can be found in the Data Classification Toolkit.

International considerations

In considering who might receive de-identified or anonymized data sets, we must consider whether those data sets will leave the country of origin. Is the data being provided to a country with different laws and regulations on privacy? Can we control where the receiver stores our data if the law restricts the transmission or storage of such data (e.g., certain research data) outside the US?

Providing services for de-identifying or anonymizing data

Many institutions provide central data anonymization services. Doing so helps bring consistency of practice and contain risk to the institution. Providing data de-identification or anonymization services at an institutional level, however, poses many challenges:

- What unit would provide the service?
- How would the unit be funded for this activity?
- What types of data will the service de-identify?
- How will the risk and value be determined?
- Who can use this service?
- Who acts as the data handler, to de-identify the data?
- How do you ensure that this data handler employs and maintains adequate safeguards?
- How is the de-identified data checked for accuracy and anonymity before providing it to the requestor?
- How do you ensure that de-identified data meets legal or regulatory requirements, if applicable?
- How do you ensure that requestor maintains the de-identified state of the data?
- What type of user support is needed?
- What are the minimum service expectations?
- What if the service does not meet expectations?
- And finally, in the end, what if the handler determines that the data cannot be de-identified or anonymized?

Top of page

General Recommendations

Organizations vary in size, type, and capacity to deal with data de-identification issues. Typically when a data de-identification need arises, information policy, security, or privacy staff assist by framing the discussion and helping find appropriate solutions. The list of recommendations included below is intended to help institutions respond to their own unique challenges regarding de-identified or anonymized data sets. This list is not intended to be complete.

Governance Recommendations

- Stewards/Stakeholders: Position the owners/stewards/stakeholders of the identified data set to take a leadership role in all decision making processes.
- Consultation: Consult with the appropriate Institutional Review Boards, data stewards, stakeholders, and subject matter experts; research compliance, HIPAA compliance, and other compliance offices; and the General Counsel's Office, Information Security Office, and Information Privacy Officer.
- Receiver agreement: Create a standard contract or service level agreement to be used with the receiver of the de-identified data.
- Due diligence: Due diligence should be conducted to determine the viability of the data de-identifier and the receiver. 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.

Process Recommendations

- Risk/benefit analysis: Identify and understand the risks and benefits of the service. Recognize that de-identification failures and re-identification efforts of receivers will potentially involve or at least reflect on the university. Honestly compare costs of providing de-identification services, including costs to manage the receiver relationship, against the benefits of the intended use of the de-identified data.
- Lower risk candidates: When considering de-identification services, ideal candidates will be those that involve information with lower risk of re-identification or that are classified into a level that requires little to no protections. These are likely to represent the best opportunities for maximizing benefit while minimizing risk.
- Higher risk candidates: Data which is questionable as to whether or not it actually can be completely de-identified (such as network flow data, web traffic data, etc.), are necessarily higher risk candidates and require careful scrutiny and consideration, and stronger strategies for reducing the risk to acceptable levels. Data classified into levels that require medium to strong protections also are higher risk candidates. Also, small data sets are generally riskier, due to the increased chances that an individual could be identified.
- Centralized de-identification services: Consider leveraging internal services when looking for ways to provide data de-identification to university community members for university purposes, e.g., create a data lab/virtual server solution, with trained data de-identification experts. Develop an institutional standard for data anonymization.
- De-identifier safeguards: Ensure the data handler doing the de-identification implements physical, technical, and administrative safeguards appropriate to the risk. Areas to explore with the de-identifier include privileged user access, regulatory compliance, data location, data segregation, recovery/data availability, personnel practices, incident response plans, and investigative/management support. Scrutinize any gaps identified.
- Proportionality of analysis/evaluation: The depth of the above analysis and evaluation and the scope of risk mitigation measures and required assurances must be proportional to the risk involved, as determined by the sensitivity level of the information involved and the criticality or value to the university of the use of the de-identified data involved. Fewer steps and strategies for mitigating risk are necessary when the risk is low; more are required when the risk is high.

Top of page

Critical First Steps

The following steps can assist in creating an institutional process for de-identifying data sets.

1. *To whom should a request for de-identified data set be made?* Determine where requests are to be directed first. Consider the data steward for that data set, or the owner of the service containing the data, or another appropriate office or group. Consider noting that any request identified as a public records request (or perhaps all external requests) go to the office that handles those requests as a first step.
2. *Who works with the requestor to understand the request, analyze the data, and identify what data elements must be de-identified?* The data steward or other receiver of the request may act as the shepherd of the request through the various steps. Consider a form or set of questions for the requestor to complete or the receiver of the request to ask the requestor; use the rest of this document to help identify what to include on your form. Document and understand the requestor's need, the existence of data to meet that need, the data elements that would need to be removed, and the risk of de-identifying and providing the data to the requestor (including the risk of re-identification). Prepare enough information to take the request to the next step. It may be necessary to include a technical expert at this stage (see step 6 for more about technical resources), in order to identify alternative solutions the requestor has not identified.
3. *From whom must approvals be obtained before the design/proposal to provide de-identified data is accepted?* Before the data set may be released? Consider setting up an approval team or process that includes the data steward, information security, information privacy, legal counsel, Institutional Review Board, and others you identify after reviewing the rest of this document. Provide the information gathered in step 2. Discuss and review the key challenges and risks. Determine most appropriate solution. Document approval, any requirements for the technical personnel who will do the de-identification, and any requirements for the requestor of the data.
4. *If there is a cost either in resources or budget to do the work, who approves the cost estimates? Who pays?* The cost of personnel to undertake the de-identification may be an issue, especially if the organization is a large research university that may receive many requests for this service. How will the institution cover these costs? Or, what charges will be made?
5. *Is it acceptable for a data steward or other members of the approval team to refuse the request on grounds other than confidentiality?* Consider allowing data stewards and/or other approval parties to refuse a project for a variety of reasons other than confidentiality, perhaps including riskiness, time, cost, resources, status of receiver, etc. as identified after reviewing the rest of this document. Who makes the final decision? Are there options that could be considered when approval parties are uncomfortable with the project as proposed? What projects receive priority?
6. *What technical resources are available to do the de-identification?* Identify technical personnel who will be trained in the issues of de-identification and the technical methods of achieving de-identification of various types of data sets. Consider limiting the number of persons performing this work in order to reduce risk to the de-identified data sets and to increase expertise in technical de-identification processes. Look to information technology, research technology, computer science, or other areas to find personnel. Provide resources to identified personnel for training and ongoing learning. Review this document with identified personnel. Consider having technical personnel sign an annual confidentiality and appropriate use agreement. Consider having information security review de-identification processes prior to implementation.
7. *Document the provision of the de-identified data to the requestor.* Identify who will provide the de-identified data to the requestor. Does the technical data handler do so, or does the handler provide the data to another party to review the data and complete the transaction? Track the provision of de-identified data to requestors, including any agreements the requestor may be made to sign.

Top of page

Further Information and Resources

Institutional Resources

- University of Chicago, Guidance on IRB Review of Research Involving Existing Data Sets
- University of Minnesota, De-Identifying Data for Research
- Yale University, Use and Disclosure of De-Identified Information and of Limited Data Sets (HIPAA Policy 5039)
- Purdue University, Network Data Collection for Research Purposes

Regulatory Resources

- Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Pub. L. No. 104-191, § 264 (1996), codified at 42 U.S.C. § 1320d; Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. § 160 (2002), 45 C.F.R. § 164 subpts. A, E (2002).
    - US Department of Health and Human Services, Understanding Health Information Privacy
    - US Department of Health and Human Services, Breach Notification Rule
- Gramm-Leach-Bliley Act (GLBA) (Pub. L. No. 106-102 (1999), privacy protections are codified at 15 USC § 6801 et seq.).
    - Federal Trade Commission, Financial Privacy
- Family Educational Rights and Privacy Act, as set forth in 20 U.S.C. §1232g (FERPA).

Technical Resources

- Internet2, Network Flow Data Privacy Policy
- Crypto-Pan is a cryptography-based sanitization tool for network trace owners to anonymize the IP addresses in their traces in a prefix-preserving manner. Note: Crypto-Pan was ported into the FLAIM project at an early stage of that effort.

Other Resources

- Center for Democracy and Technology (CDT)
    - CDT Policy Post (October 1, 2009), Stronger Protections for, and Encouraging the Use of De-Identified (and "Anonymized") Health Data
    - CDT Policy Post (October 1, 2009), Government Information, Data.gov and Privacy Implications
    - CDT, Encouraging the Use of, and Rethinking Protections for De-identified and "Anonymized" Health Data (June 2009)
    - CDT, Compendium of "Sensitive" Information Definitions (March 24, 2008)
- Computerworld, Privacy Matters: When is personal data truly de-identified? (July 24, 2009)
- Electronic Health Information Laboratory, CHEO Research Institute
- SAS Global Forum, Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification (also see the video presentation)
- U.S. Department of Education's Privacy Technical Assistance Center, PTAC Toolkit (includes an overview of data de-identification and disclosure avoidance techniques)

Top of page

Questions or comments? Contact us.