ISSN 2161-3974
Commentary

Journal of eScience Librarianship 13 (3): e965
DOI: https://doi.org/10.7191/jeslib.965

Journal of eScience Librarianship
putting the pieces together: theory and practice

Understanding how to identify and manage personal identifying information (PII) to further data interoperability

Zixin Nie, RTI International. Durham, North Carolina, USA, zixinnie@rti.org

Abstract

Respect for research participant rights is a key aspect for consideration when creating and utilizing interoperable data. From that perspective, requirements for sharing research data often call for the data to be de-identified, i.e., the removal of all personal identifying information (PII) prior to data sharing, to ensure that the participant's data privacy rights are not infringed upon. However, what constitutes PII is often a point of confusion amongst researchers who are not familiar with privacy laws and regulations. This paper hopes to provide some clarity around what makes research data identifiable by presenting it under a different perspective from what most researchers are familiar with. It also provides a framework to help researchers determine where PII could exist within their data that they can use to help with privacy impact evaluations. The goal is to empower researchers to share their data with greater confidence that the privacy rights of their research subjects have been sufficiently protected, enabling access to greater amounts of data for research use.

Received: July 15, 2024 **Accepted**: November 26, 2024 **Published**: December 13, 2024

Keywords: personal identifying information, de-identification, data interoperability, data privacy, HIPAA, HIPAA Safe Harbor, HIPAA Expert Determination, privacy evaluation

Citation: Nie, Zixin. 2024. "Understanding how to identify and manage personal identifying information (PII) to further data interoperability." *Journal of eScience Librarianship* 13 (3): e965. https://doi.org/10.7191/jeslib.965.

The *Journal of eScience Librarianship* is a peer-reviewed open access journal. © 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
See http://creativecommons.org/licenses/by/4.0.
OPEN ACCESS

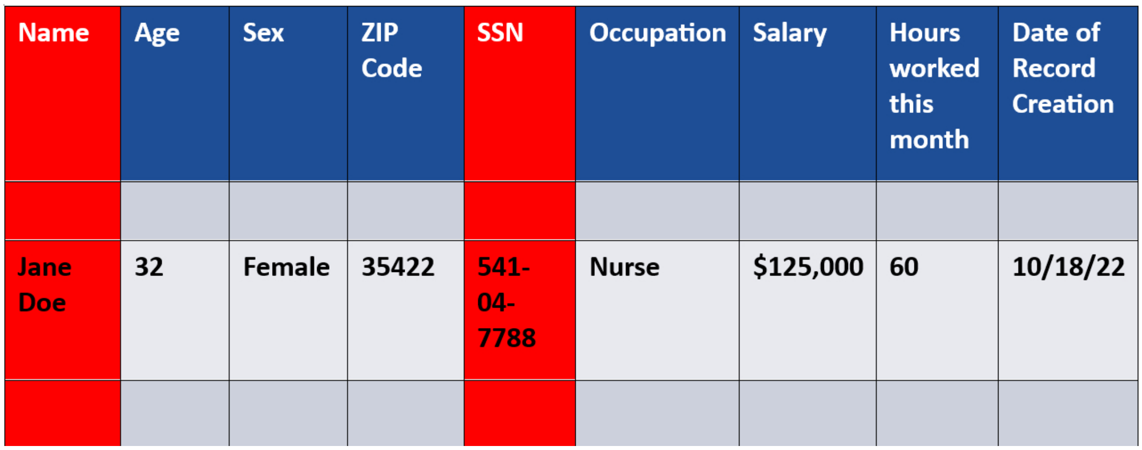
Introduction

Legal definitions and guidance for what constitutes Personal Identifiable Information (PII) can lead to misconceptions from researchers and data managers, affecting situations where data de-identification is necessary, such as when re-using or sharing data outside of explicit data subject consent. Guidance from regulators, such as the definition for PII provided by the Department of Labor, describe PII as data that can "directly identify an individual (e.g., name, address, social security number or other identifying number or code, telephone number, email address, etc.)" or data that can "identify specific individuals in conjunction with other data elements, i.e., indirect identification" (Department of Labor, n.d.). This misconception is further reinforced by rules-based de-identification methods such as the Safe Harbor provision under the Health Insurance Portability and Accessibility Act (HIPAA). Under HIPAA Safe Harbor, de-identification involves the removal of 18 different types of identifiers (Office for Civil Rights (OCR) 2012). Given that the requirements under HIPAA are the most prescriptive under US law, and that other privacy regulations do not have explicit requirements for what makes data identifying and how to de-identify data, the rules-based de-identification method defined under HIPAA Safe Harbor has become a standard for how identifiable information is conceived (Sweeney et al. 2017).

Definitions like these create misconceptions that what constitutes PII are specific fields that are named in the provided guidance, giving the impression that if those fields are removed from the data, then the data is no longer PII. Conceiving of identifiable information in such a matter is problematic for two reasons: 1) privacy regulations in other parts of the world, such as the European Union (EU), have different requirements for what data is considered identifiable and how data should be anonymized, calling for a risk-based approach (European Data Protection Supervisor 2021); and 2) this conception of identifiable information separates identifiability from the data subjects, creating situations where the data may be considered de-identified, but data subjects could still be identifiable. A shift in the way PII is conceived and described can help researchers and data managers better identify and protect PII they manage. This paper describes a different way to think about PII to facilitate this shift and provides a framework to help researchers and data managers identify PII in the data that they hold.

A Change in Thinking about PII

Legal definitions and regulatory guidance make it seem like what makes data identifiable are the presence of certain fields within your data. This type of thinking is illustrated in Figure 1, where the presence of the highlighted columns "Name" and "SSN" are what makes the data identifiable. However, if this framework of thinking is used, there is unclarity around "indirect identification", i.e., one could potentially use the fields "Age," "Sex," "ZIP Code," "Occupation," "Salary," "Hours worked this month," and "Date of record creation" to identify data subjects as well, depending on what other information about the data subjects is available. Does this mean that all the other fields contained within this dataset are PII as well? Thinking about PII in this way leads to unclarity as to what constitutes PII, which could result in either "under de-identification", in which

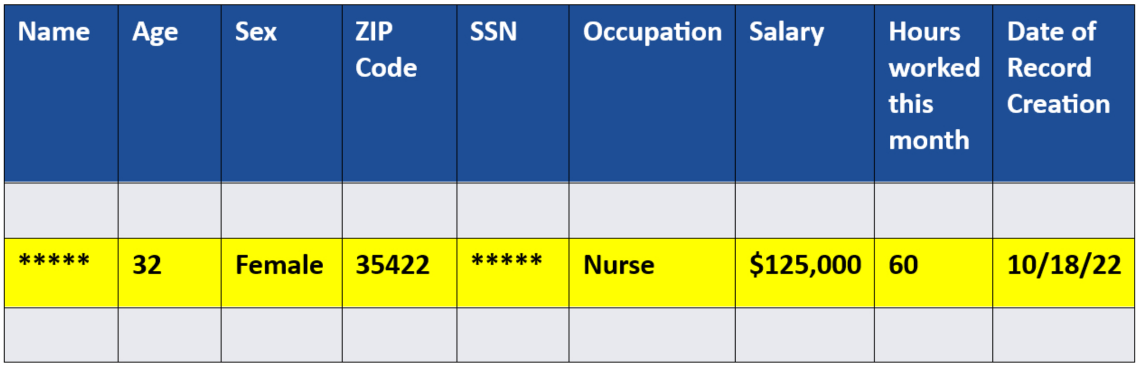


Name	Age	Sex	ZIP Code	SSN	Occupation	Salary	Hours worked this month	Date of Record Creation
Jane Doe	32	Female	35422	541-04-7788	Nurse	$125,000	60	10/18/22

Figure 1: Illustration of "field-level" thinking about PII. In this figure, the directly identifying fields of "Name" and "SSN" are highlighted. However, within Jane Doe's record, there may remain quasi-identifying information that could lead to her re-identification even once the directly identifying fields are entirely removed (i.e., column suppression) from the data.

there is still enough information left within the data to re-identify data subjects, or "over de-identification," in which so much information about data subjects gets removed that there are negative impacts upon data utility.

Instead of thinking about PII in terms of what fields exist in the data, one should think about PII in terms of what records could be identifiable based on their contents. This kind of thinking is illustrated in figure 2, where the focus is on the rows (i.e., records), instead of the columns (i.e., fields). In the example record, the contents of the "Name" field and the "SSN" field lead to the direct identification of the data subject, making the whole record PII. Removal of the contents of the "Name" and "SSN" fields constitute de-identification of the record; however, if the remaining information in the record is sufficient to re-identify the data subject (either through linkage to other data or through other external knowledge) the record remains PII. Ensuring that data is not PII means proving that the contents of records cannot be used in conjunction with external knowledge to identify data subjects, which may require a combination of statistical risk measurement as well as assessment of privacy, security, legal, and contractual controls.



Name	Age	Sex	ZIP Code	SSN	Occupation	Salary	Hours worked this month	Date of Record Creation
*****	32	Female	35422	*****	Nurse	$125,000	60	10/18/22

Figure 2: Illustration of "record-level" thinking about PII. Jane Doe's record (the data row), not the fields (data columns) is what contains identifying information about Jane Doe. We can use this kind of thinking to adjust the specific values of fields within her record to de-identify her, instead of applying blanket actions across entire fields. This also enables us to measure the likelihood that Jane Doe can be re-identified based on the information in her record.

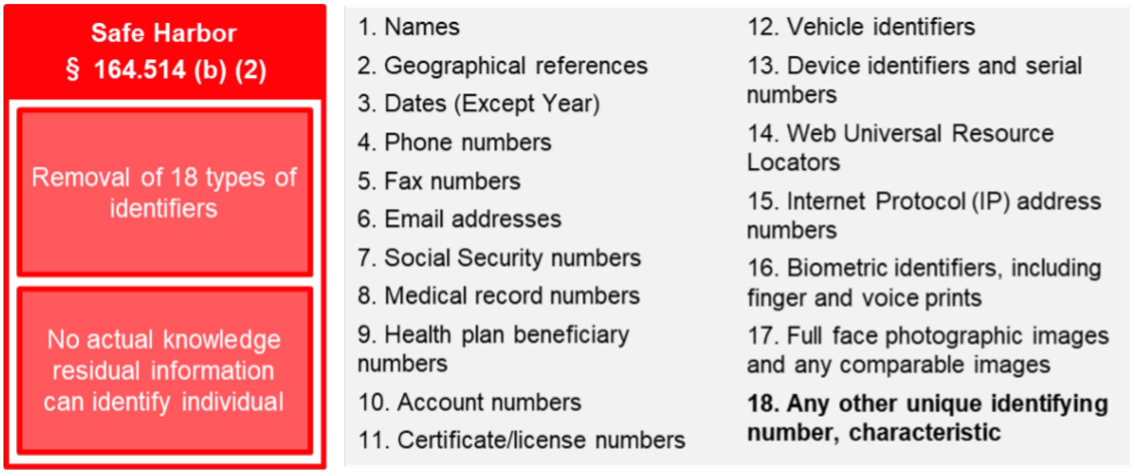
Frameworks for Identifying PII

Frameworks for determining whether data contains PII come in two methods. The first is rules-based, best exemplified by HIPAA Safe Harbor, where data is PII if it contains any one of 18 types of information (detailed in figure 3). Such methods are simple to understand and provide prescriptive guidance for de-identification (remove the 18 types of identifiers), however they can lead to over or under de-identification of data. The second method is to use a classification framework to determine what record contents could contain information that could lead to identification of a data subject, and then conduct record-level risk measurement. This section will provide details about the second method, providing guidance on how it can be utilized.

Classification Framework

Classification involves grouping record contents into three categories: direct identifiers (DIs), quasi-identifiers (QIs), and Non-Identifiers (NIs). DIs are sufficient to identify a data subject by themselves, either by providing the data subject's identity or by providing a code that can link to other data that contains the data subject's identity. These include names, social security numbers, home addresses, email addresses, phone numbers, medical record numbers, credit card numbers, and license plate numbers. QIs can be used in conjunction with other QIs to identify a data subject. These include sex, age, race and ethnicity, ZIP code of residence, occupation, medical diagnoses, income, education, and dates of events. Non-identifiers are variables whose contents cannot be used to identify a data subject.

For record contents to be identifying, they need to satisfy the conditions of being replicable, distinguishable, and knowable.



Safe Harbor
§ 164.514 (b) (2)

Removal of 18 types of identifiers

No actual knowledge residual information can identify individual

1. Names
2. Geographical references
3. Dates (Except Year)
4. Phone numbers
5. Fax numbers
6. Email addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators
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**

Figure 3: HIPAA Safe Harbor de-identification requirements. HIPAA Safe Harbor is a prescriptive, rules-based framework that calls for de-identification of Personal Health Information through the removal of 18 types of identifiers, as well as no actual knowledge that the information could be used alone or in combination with other information to identify a data subject (Office for Civil Rights (OCR) 2012).

- **Replicable** means that the content does not change over time. For example, the sex of a data subject generally does not change after the subject's birth, so it is replicable. In contrast, a data subject's blood pressure can fluctuate with each measurement, making it not replicable.
- **Distinguishable** means that the content can be used to tell apart data subjects. For example, sex can be used to distinguish between data subjects that are males and females. It is important to note that whether a variable remains distinguishable depends on the population the data is extracted from, i.e., in a dataset of testicular cancer patients, sex ceases to be distinguishable, as all data subjects should be male.
- **Knowable** means that the content is something that someone using the data can know about the data subject. Determining knowability requires understanding the context surrounding the data (i.e., intended use, access and security controls, and contractual obligations) and modelling different actors who could identify data subjects. Broadly speaking, knowability can be broken down into two levels:
    - **Publicly knowable**: information that can be obtained from publicly available sources, such as voter registries, newspapers, obituaries, and social media. These include demographic information such as age, sex, date of birth and death, and newsworthy events like vehicular accidents or crimes.

- **Acquaintance knowable**: information that could be known by acquaintances of data subjects, which can include friends and family, coworkers, classmates, employees, and social media followers. These can include medical diagnoses, income, financial transactions, and important events. With the advent of social media and the increased public sharing of information online, information that was considered knowable only to acquaintances can become publicly knowable if shared without proper controls (i.e., making photos from an event visible to the public instead of visible to just friends).

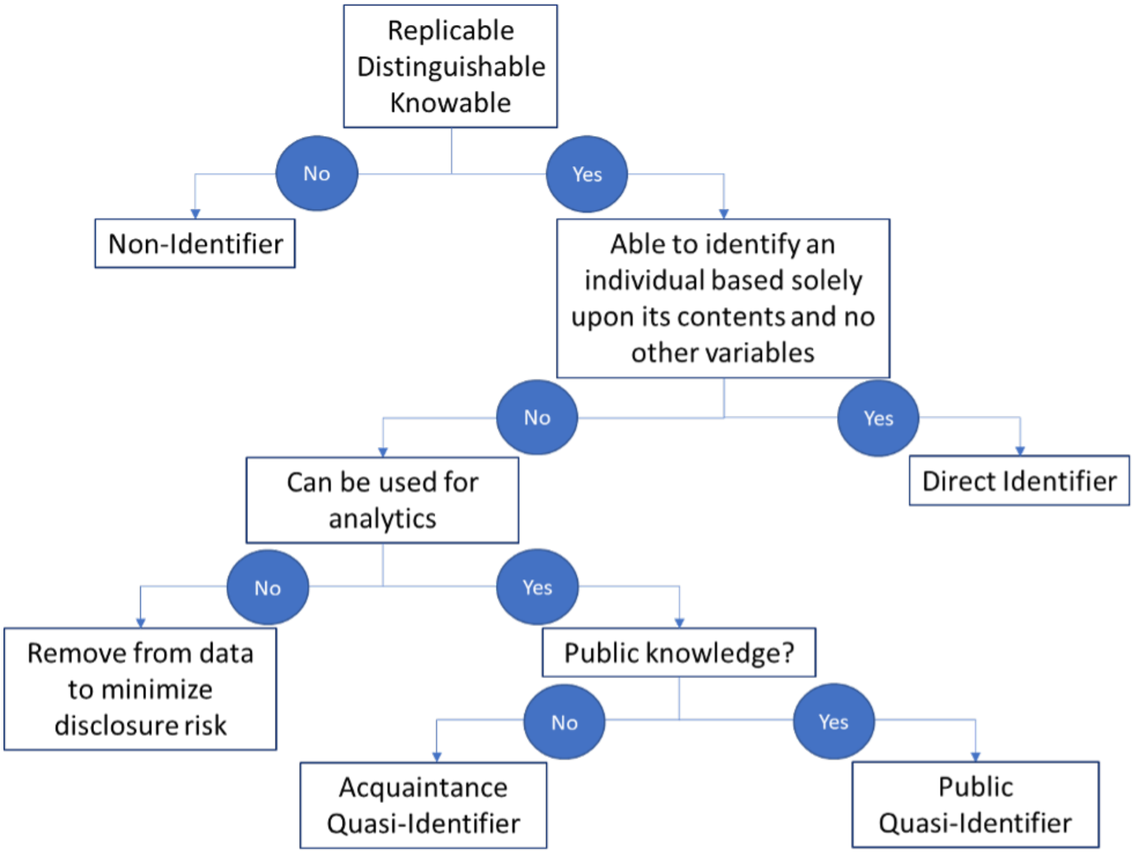

Replicable
Distinguishable
Knowable
No
Yes
Non-Identifier
Able to identify an individual based solely upon its contents and no other variables
No
Yes
Can be used for analytics
Direct Identifier
No
Yes
Remove from data to minimize disclosure risk
Public knowledge?
No
Yes
Acquaintance Quasi-Identifier
Public Quasi-Identifier

Figure 4: Identifier classification flow-chart. Researchers can use this flow-chart to determine where potential identifying information could lie within their data records.

Figure 4 provides a flowchart which can be used to help guide classification efforts. The steps in this process are as follows:

1. Determine the information within the data that satisfies the conditions of being replicable, distinguishable, and knowable;
2. Break that information down into DIs and QIs;
3. Separate QIs into three buckets, publicly knowable, acquaintance knowable, and not useful for analytics;
4. For data minimization purposes, it is recommended that QIs that are not useful for analytics are removed from the dataset, which will also reduce data subjects' identification risk.

Risk Measurement Methodology

After performing classification, the next step is to measure the likelihood (or risk) that a data subject can be identified based on the information within their records. There are many methodologies for calculating this, including using statistical estimation methods such as the Zayatz estimator for population-level uniqueness (Zayatz 1991), K-anonymity and K-mapping (El Emam and Dankar 2008, Pannekoek 1999), and information theory (Zhang, Lu, and Tian 2019), as well as methods that evaluate external identifiable information about data subjects known as motivated intruder tests (Information Commissioner's Office 2021). These methods may require specialized expertise to be conducted, which may preclude researchers from performing these measurements themselves. There are some guidelines researchers can use to help them determine whether records within their data could be PII:

1. Records contain direct identifiers should be considered PII;
2. Records that contain numerous publicly knowable QIs are at high risk of being PII. A study conducted in 2000 found that 87% of the United States population can be uniquely identified based on 5-digit ZIP code, gender, and date of birth (Sweeney 2000). Since then, even more information has entered the public domain, the risk that someone can be identified based on their publicly knowable information has increased;
3. If records have been de-identified so that only acquaintance knowable QIs are present, it is likely that they would not be PII. However, researchers should still be careful as to who has access to the data to prevent inadvertent re-identification by acquaintances of data subjects.

It is important to note that measuring re-identification risk involves evaluating the contents of datasets and the context surrounding the data. This means evaluating the privacy and security controls of the organizations and environments that hold data and determining motivations and capacity for identification of data subjects. A framework for conducting these kinds of evaluations has been detailed in an international standard, ISO/IEC 27559: 2022 (International Organization for Standardization 2022).

Conclusion

The traditional way of thinking about PII in terms of the fields contained within a dataset creates misconceptions about what is identifying information, creating situations where datasets may be de-identified improperly, causing risks to data subject privacy. Considering PII in terms of records instead of fields helps to mitigate these risks. The classification framework and the guidelines presented in this paper can give researchers an idea of where re-identification risks may lie within their data. For an accurate picture, experts can measure the likelihood of identification of each record, providing targeted mitigation strategies to protect data subject privacy while maintaining data utility.

Competing Interests

The author declares that they have no competing interests.

Acknowledgements

The content of this article is based on the presentation entitled "Understanding how to identify and manage personal identifying information (PII) to further data interoperability" originally presented at RDAP Summit 2024, available from Open Science Framework: https://osf.io/ybhe8.

References

Department of Labor. n.d. "Guidance on the Protection of Personal Identifiable Information." DOL. Accessed June 10, 2024. https://www.dol.gov/general/ppii.

El Emam, Khaled, and Fida Kamal Dankar. 2008. "Protecting Privacy Using K-Anonymity." Journal of the *American Medical Informatics Association: JAMIA* 15 (5): 627–637. https://doi.org/10.1197/jamia.M2716.

European Data Protection Supervisor. 2021. "10 Misunderstandings Related to Anonymization." https://www.edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf.

Information Commissioner's Office. 2021. "Anonymisation: Managing Data Protection Risk Code of Practice."

International Organization for Standardization. 2022. "Information Security, Cybersecurity and Privacy Protection – Privacy Enhancing Data de-Identification Framework." ISO/IEC Standard No. 27559:2022. https://www.iso.org/standard/71677.html.

Office for Civil Rights (OCR). 2012. "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." Text. HHS.Gov. September 7, 2012. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Pannekoek, Jeroen. 1999. "Statistical Methods for Some Simple Disclosure Limitation Rules." *Statistica Neerlandica* 53 (1): 55–67. https://doi.org/10.1111/1467-9574.00097.

Sweeney, Latanya. 2000. "Simple Demographics Often Identify People Uniquely." Carnegie Mellon University. Data Privacy Working Paper 3. Pittsburgh. https://dataprivacylab.org/projects/identifiability/paper1.pdf.

Sweeney, Latanya, Ji Su Yoo, Laura Perovich, Katherine E. Boronow, Phil Brown, and Julia Green Brody. 2017. "Re-Identification Risks in HIPAA Safe Harbor Data: A Study of Data from One Environmental Health Study." Technology Science 2017:2017082801.

Zayatz, Laura Voshell. 1991. "Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample." *Proceedings of the Section on Survey Research Methods*, August. https://www.census.gov/content/dam/Census/library/working-papers/1991/adrm/rr91-08.pdf.

Zhang, Zeyu, Zhiyang Lu, and Youliang Tian. 2019. "Data Privacy Quantification and De-Identification Model Based on Information Theory." In *2019 International Conference on Networking and Network Applications (NaNA)*, 213–222. https://doi.org/10.1109/NaNA.2019.00046.