



June  
2025

# Testing Privacy-Preserving Record Linkage Within a Statewide Longitudinal Data System: *Lessons Learned*

---

Amy O'Hara, Ph.D, Research Professor

Stephanie Straus, M.Ed, Policy Fellow

Justin Liu, Research Assistant

Massive Data Institute,  
Georgetown University,  
McCourt School of Public Policy

**GEORGETOWN**  
**UNIVERSITY**  
McCourt School of Public Policy

**MASSIVE**  
**DATA**  
**INSTITUTE**

## ACKNOWLEDGEMENTS

This report was funded by the Gates Foundation. The views expressed are those of the author(s) and should not be attributed to the foundation.

We would also like to thank the teams at the Department of Transformation and Shared Services, Arkansas Data Office (ARData)—Robert McGough, Veronika Gudipati, and Nishav Mainali—and at Medical Research Analytics and Informatics Alliance—Kruti Doshi and William Trick—for their collaboration and contributions to this project

# Contents

<b>Introduction</b> .....	001
<b>Use Case</b> .....	002
<b>Interest in PPRL</b>	
<b>Collaborators</b>	
MRAIA	
<b>Pilot Implementation</b>	
ARData's Data Goals	
Implementation	
Evaluation	
<b>Lessons Learned</b> .....	006
<b>What Worked</b>	
<b>Other Considerations</b>	
<b>Next Steps</b> .....	008
<b>References</b> .....	009
<b>Appendix</b> .....	009



# Introduction

Linking data across silos is a key feature of integrated data systems. Specifically, [Statewide Longitudinal Data Systems](#) (SLDS) often need to link education datasets to social services, health, and workforce-related datasets to obtain a holistic picture of their student populations. In order to link the data from these fields, which are housed in disparate systems and agencies, students' Personally Identifiable Information (PII) must be transported into additional data systems and accessed by additional staff, which increases the risk of unauthorized data access, unintentional disclosures, and hacks or breaches. This project demonstrates how a privacy-enhancing technology (PET) called privacy-preserving record linkage (PPRL) can securely link data across siloed systems without ever disclosing PII, thereby maintaining data utility without compromising privacy.

Privacy-preserving record linkage (PPRL) is a suite of techniques that links data across separate datasets without disclosing any sensitive information within them (Gkoulalas-Divanis, Vatsalan, Karapiperis, & Kantarcioglu, 2021). Since PII is required for locating individuals across two or more data files, PPRL leverages various encryption algorithms to perform the record linkage process on obfuscated versions of the PII.

PPRL can enhance operational, research, and administrative practices, such as:

- Safe production of statistics required for mandatory reporting
- Delivery of support services for constituents
- Evaluation of programs and policies, internally or via external researchers

SLDS staff need to be able to securely link sensitive data across silos and grant access to this linked data to run their day-to-day operations. PPRL is a useful tool in that it allows for the robust linking of disparate datasets without ever exposing the original files' PII. One of the major mechanisms used in PPRL to 'obfuscate' PII is called hashing, which is the replacement of sensitive inputs with a random string of characters (hash) unique to each original record in the data. Hashes are designed so that they cannot be reversed, and many hashing algorithms have been [approved](#) by the National Institute of Standards and Technology.

Hashing-based PPRL has been empirically used in the government and nonprofit spheres, especially in healthcare. One of the leading examples of large-scale PPRL is [CAPriCORN](#), a Chicago-based research network that joins over 12 million data records across medical institutions and community partners in order to improve patient health outcomes. CAPriCORN leverages a PPRL software that cleans and then hashes data across healthcare institutions (Kho et al., 2015). An independent third party links the individual hashed records to reconcile the patients across the various datasets.

Within education, PPRL has been used to securely link education data to related fields. Researchers at the Georgia Policy Lab, with support from the Administration for Children and Families, created a secure hashing tool called [Solution for Identifying Linkage Keys \(SILK\)](#), which hashes social security numbers to allow for privacy-preserving linkages across state departments of education and social and family services. The open-source SILK tool allows the government partners to examine family service data securely, ensuring confidentiality while fostering collaborative insights.

The California Policy Lab utilized PPRL in their [CalFresh](#) study to link data in partnership with community colleges, universities, Free Application for Federal Student Aid (FAFSA), and Supplemental Nutrition Assistance Program (SNAP) to examine food insecurity among California's college students. This secure linkage enabled a comprehensive analysis of student participation in CalFresh food benefits, revealing barriers faced by students while protecting their privacy. Insights from this PPRL-enabled study are helping inform strategies to increase access to CalFresh benefits for students in need across California.

Building on these PPRL use cases, this paper outlines the pilot testing of PPRL using state workforce and public education data through a collaboration with the Arkansas Department of Transformation and Shared Services, the Medical Research Analytics and Informatics Alliance (a health information technology nonprofit), and Georgetown University.

## Use Case

[Arkansas Data \(ARData\)](#) is a collaborative effort within the [Arkansas Department of Transformation and Shared Services](#) (TSS) that works across state agencies. TSS was established in 2019 for the purpose of creating a more efficient government through service delivery and collaboration across state governments. To that end, ARData maintains a high-functioning SLDS in order to use data-driven decision-making processes for continuous improvement and routine problem-solving processes.

### Interest in PPRL

ARData staff researchers are interested in designing and testing PPRL approaches across their education and workforce data. PPRL can help alleviate some of their privacy concerns by reducing the sharing of PII across Arkansas's SLDS agencies.

The ARData team currently creates linkages across SLDS program files using exact and fuzzy matching techniques on data under their control. Exact, or deterministic, matches are often used when ARData is linking between two datasets containing a unique identifier, such as linking postsecondary education data to Unemployment Insurance wage data using SSN. Exact matching requires that a PII data field, such as an SSN, corresponds exactly across two or more datasets in order for those two entities to be considered a match. For datasets that do not contain unique

identifiers, though, such as K-12 education data, ARData uses a combination of exact and fuzzy (inexact) matching, with various PII variables, such as first name, last name, and date of birth. Fuzzy matching allows for 'inexact' matching in that it permits slight differences in the PII data fields across two or more datasets—for example, typos or transpositions of characters in a last name—and can still consider those two entities to be a match.

ARData is interested in testing PPRL on Arkansas postsecondary education and workforce data, replicating the aforementioned linkages already done 'in the clear,' or without any formal privacy protection. These linkages allow for the production of important statistics on the state of their constituents, such as these [higher-education-through-employment dashboards](#).

Other use cases for ARData PPRL can include:

- Improving upon their existing linkage bias by exploring the different PPRL matching methods, and comparing the results with their own;
- Facilitating matches beyond common SLDS datasets, such as linking AR's National Guard data with their state UI wage data to explore military outcomes; and
- Enhancing grant reporting requirements, such as compiling employment statistics on enrollees of Department of Labor-funded training programs.

The value of utilizing PPRL within an SLDS rests largely in the risk and utility calibration that this privacy-preserving method affords: state agencies and departments can retain or improve the existing utility of their datasets, while still adequately preserving the privacy of the students and individuals housed within these datasets. Agencies can share record-level data with each other or externally without having to disclose sensitive PII fields. PPRL is a valuable stepping stone for any integrated data system working towards a tiered access model in the future.

## Collaborators

The Massive Data Institute (MDI) at Georgetown University, an interdisciplinary research institute housed in the McCourt School of Public Policy, approached ARData in the winter of 2024 to discuss a demonstration project involving PETs. ARData expressed interest in PPRL. This led to a research partnership between ARData and MDI, which expanded to include MRAIA as the Technical Assistance provider. Together, the three organizations have mapped out the testing of PPRL on select ARData datasets, addressed administrative and regulatory considerations for this pilot, and considered additional PPRL needs for future projects, using the tools and training provided by MRAIA.

### MRAIA

The Medical Research Analytics and Informatics Alliance, better known as MRAIA, is a not-for-profit, mission-driven organization created to serve the immediate and growing demand for health information technology solutions. Founded in 2011 by a team of physicians, researchers, and leaders in medical and public health informatics, MRAIA serves as an honest data broker by securely

housing, managing, integrating, and translating data into meaningful information that can be used to improve lives. MRAIA works with partners across a variety of sectors (government, academia, healthcare, and corporations) to enable collaboration and synergy.

MRAIA brings PPRL expertise to this project through their open-source, record linkage software system called [Linkja](#). Linkja is designed to link individuals across disparate datasets from separate institutions or agencies while protecting their privacy through a de-identification process that occurs at each institution. Linkja was developed as a collaboration between academic and public health system developers and researchers.

In order to prepare the ARData team to effectively utilize their PPRL product, MRAIA trained them on Linkja, which hashes combinations of PII elements using a standard Secure Hashing Algorithm, and then offers a menu of linkage rules to provide options to maximize accuracy. MRAIA provided the [Linkja documentation site](#), which contains links to the necessary software dependencies and to the code modules in their GitHub repository. ARData staff installed the components locally on their machines. Only ARData accessed their own, confidential data with which to test Linkja; MRAIA provided technical guidance on installation, implementation, and troubleshooting, without ever seeing ARData data.

## Pilot Implementation

### ARData's Data Goals

ARData has a complex master data management (MDM) system that reconciles many instances of the same individual across various datasets in the SLDS. To begin their PPRL testing, ARData used an existing synthetic dataset generated by The Center for Advanced Research in Entity Resolution and Information Quality (ERIQ), which is based in the College of Engineering at the University of Arkansas. The synthetic dataset was generated to represent their MDM system, and to reflect their SLDS datasets, such as K-12, postsecondary, and workforce data with varying PII identifiers. For testing Linkja, ARData utilized a synthetic dataset with 3,000 records containing First Name, Last Name, Social Security Number (SSN), and imputed Date of Birth (DOB).

## Implementation

Linkja inputs four PII variables for its PPRL algorithm (items 1-4); a fifth input variable is inputted and included in the local output file solely for local data validation activities:

1. First Name (2+ characters required for processing)
2. Last Name (2+ characters required for processing)
3. Date of Birth (DOB)
4. Social Security Number (SSN) (optional)
5. Unique [student] ID (for local data validation)

All variables except SSN are required in the datasets to run Linkja. Linkja cleans input variables to prepare them for linkage, such as truncating SSN into the last 4 digits only. Linkja hashes allow for 10 Rules, or linkage methods, for projects to select; each Rule uses a different combination of the four PII variables. Each of the 10 possible matching rules is performed by hashing the various combinations of PII into discrete tokens, and then performing deterministic and rule-based fuzzy matching on those tokens. Linkja's combinations of PII variables and the 10 Rules are based on a detailed manual matching validation process using health system data obtained from difficult-to-match populations. MRAIA built important data quality features into Linkja to account for common record linkage errors and data quality issues that can reduce matching accuracy (Trick, Doshi, Ray, & Angulo, 2019). For example, Linkja cleans the PII variables for removal of special characters (e.g., hyphens and apostrophes), scans them for data entry errors, and allows for common 'transpositions,' such as the flipping of characters within a field (e.g., last name first name) and common data entry keystroke errors such as 1- or 365-day offsets for DOB.

The ARData team prepared their input data—the synthetic dataset—to meet Linkja's PII requirements. They tested all components of the system, including generation of the cryptographic keys and files, the hashing, and the linking, testing seven of the ten Linkja matching Rules.

## Evaluation

To evaluate Linkja's performance on the ARData synthetic data, ARData examined several key results. First, the number of records in their synthetic dataset with the necessary data specifications was around 2,700 out of 3,000 original records. Approximately 300 records were not processed into hashes because they did not meet the required criteria for inputting PII variables (see Implementation).

The ARData team tested Linkja [Rules 3-10](#) (See Table 1) on the 2,700 records. The matching combination, or Rule, that optimized the linkage accuracy was Rule 3 (Full Name, DOB, and SSN). This means that the synthetic dataset representing, say, AR postsecondary data matched with the highest level of overall accuracy to the synthetic dataset representing, say, UI wage data, when the two separate datasets had all four PII fields (full First Names, Last Names, DOB, and SSN fields).

In testing their synthetic data without PPRL, ARData expects ~1,700 matches. With Linkja, ARData received ~1,800 matches with its most effective matching combination. The accuracy of this matching combination, or the agreement between ARData and Linkja's matching protocols, was 88% (see Appendix for further details). This means that, for 88% of the time, Linkja declared two records to be a match, or to be a non-match, in the same way that ARData did outside of PPRL.

# Lessons Learned

## What Worked

- **Ease of Linkja Installation and Use.** Linkja's [website](#) comes with clear [documentation](#) on the exact steps any user must take in order to utilize the tool. It contains specific instructions for downloading software dependencies and supporting programs, and code for running the three PPRL modules. The Linkja suite of tools can be run on local computers and does not require any additional infrastructure, like an isolated server. Processing time is minimal, with the ARData team logging less than 1 minute per PPRL 'Rule.'
- **Linkja's Certification and Reputation.** Linkja's reputable citations across government health agencies likely boosted its security review at ARData, where its federal government use cases such as those through the National Institutes of Health lent credence for security information officers who were seeing this tool for the first time (Boulger, Hinami, Lyons, & Nowinski, 2022; Trick et al., 2021). Further, other SLDS may find it useful that Linkja has been certified to meet HIPAA requirements for disclosure of health data. Linkja was reviewed and certified by cryptography experts Dr. Fritz Scheuren and Dr. Patrick Baier as meeting HIPAA's rules for de-identification. Such certification helps government staff gain stamps of approval from their Institutional Review Boards and legal and compliance teams for data linkage needs.
- **Linkja takes into account common data quality issues that can lead to linkage errors.** SLDS data, by nature of its integration across various departments, can contain errors, typos, and intentional corruption in their PII variables, which makes data linkage efforts very challenging. Linkja combats this through several checks and features built into its coding modules. For example, Linkja tests the validity of SSN fields before hashing them, eliminating invalid SSNs such as those that are all 9's or all 0's. Further, Linkja accounts for certain typos and the miskeying of PII fields, such as a transposed full name (Last Name + First Name instead of First Name + Last Name) and 1- or 365-day offsets for DOB (1995-01-17 instead of 1994-01-17). These data quality checks built into Linkja help increase the chance of more successful matches.

## Other Considerations

- **Linkja Approval and Implementation Assistance.** Even though Linkja is open-source, the tool still had to undergo a standard security review by Arkansas information security officers. Once the software and code were whitelisted with their security team, ARData needed their IT support team to grant them administrator rights in order to begin installation. Additionally, programming language expertise varies by SLDS, but in the case of ARData, their data scientists were familiar with Python, not Java (one of Linkja's major languages). It was key to have MRAIA as a Technical Assistance provider to offer troubleshooting help to ARData as they ran the coding modules.

- **Linkja’s Required Attributes.** Linkja requires that certain PII variables, such as Date of Birth (DOB), are present in the datasets. However, certain SLDS datasets do not have all of the required variables—Arkansas’ UI wage data, for example, only recently added DOB. The ARData team needed to impute DOB for each record, since Linkja will not run if any of its required attributes are missing. For broad use in across-SLDS linkages, it would be helpful to have a ‘feature flag,’ or a configuration file that tells software how to operate, that gives the option to toggle certain data schema requirements on and off, instead of having to change the codebase.
- **Linkja only ‘fuzzies’ parts of the matching processes.** Linkja has exact, or deterministic, matching capabilities built into its Rules. The Rules allow for some ‘fuzzing’ of PII elements themselves (the common typos and transpositions), which are then hashed, but these hashes are matched in an exact manner with no probabilistic matching. Probabilistic matching assigns weights to PII elements and then calculates the likelihood of a match between two records. It is often needed in SLDS to account for differing data quality levels across systems. In ARData’s case, probabilistic matching is useful for their SLDS data subjects without SSNs, such as [Foreign Born STEM Engineers](#) and K-12 public school students. Thus, incorporating probabilistic matching tools, such as distance measures, into Linkja could help increase match rates.

## Next Steps

ARData’s testing of Linkja showed that an open-source PPRL tool can successfully be adopted by an SLDS to facilitate asynchronous, privacy-preserved matches across data silos. Linkja is particularly useful when no trusted intermediary needs to see the data, and when exact identifiers, such as SSNs, and high-quality supporting identifiers, are present.

Next steps for ARData’s testing of Linkja include:

- **Test Linkja with real-world field studies.** The synthetic data that was used contained imputed Date of Birth, so the possible data quality errors for that PII field were greatly underestimated. Testing Linkja with K-12 or postsecondary data and UI wage data would provide a more robust assessment of Linkja’s sensitivity to the typos and permutations often seen in PII linking fields. Further, over 90% of the synthetic data records used contained SSNs. Testing Linkja on K-12 data, or any other dataset without a reliable SSN field, would help determine the tool’s ability to link SLDS data subjects without unique identifiers.
- **Test Linkja with several, distrusting parties.** ARData, for the purposes of this demonstration, assumed all three roles needed to implement Linkja: the generator of the hashes, the data contributors, and the honest broker that links the datasets together via the hashes. In the future, ARData could test the tool with two or more SLDS agencies as the data contributors, and keep themselves as the honest broker. This could reflect a more realistic governance structure, in which two or more AR state agencies both contain highly sensitive information and do not want each other, nor their honest broker, to see their PII.

- **Incorporate Linkja with outside programs beyond ARData.** ARData participates in multistate collaboratives and other external data linking efforts beyond its internal servers and networks. It could be interesting to test Linkja with these external groups, such as the [Secure Query System](#) or the [Administrative Data Research Facility](#), to see if the privacy protection of the underlying data subjects can be increased while simultaneously improving these efforts' existing match quality.

As ARData continues to refine their Linkja testing, we hope that the ARData-MRAIA partnership can serve as a model for other SLDS to test PPRL in their own systems. While complications naturally arise in implementing a new PET into an integrated data system, an adaptable, open-source solution (and solution developers) makes the troubleshooting and adjustment processes much easier. Further, the additional levels of privacy protection afforded by PPRL and other PETs allow for government staff, researchers, and evaluators to unlock insights that education-to-workforce data systems need in order to help students and workers across their states.

## References

- Boulger, J. K., Hinami, K., Lyons, T., & Nowinski Konchak, J. (2022). Prevalence and risk factors for opioid related mortality among probation clients in an American city. *Journal of substance abuse treatment*, 137, 108712. <https://doi.org/10.1016/j.jsat.2021.108712>
- Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern Privacy-Preserving Record Linkage Techniques: An Overview. *IEEE Transactions on Information Forensics and Security*. 16:4966–87.
- Kho, A.N., Cashy, J.P., Jackson, K.L., Pah, A.R., Goel, S., & Boehnke J., et al. (2015). Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *Journal of the American Medical Informatics Association*. 22(5):1072–80.
- Trick, W.E., Doshi, K., Ray, M.J., & Angulo F. (2019). Development and evaluation of record linkage rules in a safety-net health system serving disadvantaged communities. *ACI Open*. 3(02):e63-70.
- Trick, W. E., Hill, J. C., Toepfer, P., Rachman, F., Horwitz, B., & Kho, A. (2021). Joining Health Care and Homeless Data Systems Using Privacy-Preserving Record-Linkage Software. *American journal of public health*, 111(8), 1400–1403. <https://doi.org/10.2105/AJPH.2021.306304>

# Appendix

AR Synthetic Dataset			
	Matches	Unmatched	Total
Linkja	1638 (True Positives)	172 (False Positives)	1810 (matches)
	172 (False Negatives)	784 (True Negatives)	956 (unmatched)
Total	1810	956	2766 (all possible records)

## Accuracy

$(TP + TN)/2766$	$(1638 + 784)/2766$	0.88 = 88%
------------------	---------------------	------------

## Sensitivity

TP/Matches	1638/1810	0.90 = 90%
------------	-----------	------------

## Specificity

TN/Unmatched	784/956	0.82 = 82%
--------------	---------	------------