International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.13, June 2025

Privacy-Preserving Data Integration for Recidivism Assessment

Lisa Trigiante, Domenico Beneventano and Sonia Bergamaschi

Department of Engineering "Enzo Ferrari"

Università di Modena e Reggio Emilia

Via P. Vivarelli n.10, I - 41125 MODENA

ABSTRACT

The emergence of Digital Justice in conjunction with advanced Data Analysis techniques presents the opportunity to advance the criminal justice system toward an innovative Data-Driven approach. An important issue of public safety is the analysis of legal recidivism. Assessing recidivism is a complex measurement problem that necessitates reconstructing a subject's criminal history from criminal records, which usually reside in different autonomous databases. In addition, the collection and processing of sensitive legal-related data about individuals imposes consideration of privacy legislation and confidentiality implications. This paper presents the design and development of a Proof of Concept (PoC) for a Privacy-Preserving Data Integration (PPDI) framework to establish a Data Warehouse across criminal and court sources within the Italian Justice Domain and a Data Mart to assess the recidivism phenomena.

General Terms

Security, Data Management

Keywords

Data Integration, GDPR, Privacy, Pseudonym,Justice Domain

1. INTRODUCTION

The digital transformation of the Justice domain and the resulting availability of vast amounts of data describing people and their criminal behaviors offer significant promise to feed multiple research areas and enhance the criminal justice system. The recidivism phenomenon illustrates this concept as it is fundamental in criminal justice to identify the cost-effectiveness of institutional programs and prisons. *Recidivism* is a tendency of an offender to lapse into a previous pattern of criminal behavior after he has received sanctions or intervention. An important connection exists between the concept of recidivism and the growing body of research on criminal desistance. Desistance refers to the process by which a person arrives at a permanent state of nonoffending. In effect, an individual released from prison will either recidivate or desist. The statistical analysis of legal recidivism can be carried out based on data from criminal records. These records can include a wide range of data about individuals, from the basic names, ages, and addresses, to more detailed such as past addresses, relationships, and any property. These records also contain the history of a person's legal troubles, including crimes, arrests, and court cases. However, criminal records are usually distributed in different autonomous databases; for example concerning geographic or temporal criteria: the legal data of minors are separated from those of adults. Thus, each source may contain only a portion of the data regarding an individual and related sanctions.

Recidivism analysis requires a process to perform *Data Integration* (DI) and reconstruct a subject's criminal history from different autonomous criminal records. The information on a criminal record varies by country and by state, however provides a great amount of personal information; this implies the necessity to process them considering the issue of privacy and the legislation in force in the country of origin. Our research is part of a project to create a "Recidivism Data Mart" (RDM) that integrates criminal and court records within the Italian Justice Domain to enable detailed data analysis of the recidivism phenomenon that is not possible on any of the individual sources.

Compared to the conference article, the content has been revised and expanded, and a description of the PoC has been added.

This article is organized as follows. Section 2 first summarizes the three basic steps of a standard *Data Integration* (DI) process. It then defines the privacy requirements dependent on the processing of Italian personal data, established by the European *General Data Protection Regulation* (GDPR).

Extending the subject to define the concept of *Privacy-Preserving Data Integration* (PPDI) by discussing how the traditional DI steps must be adapted to prevent the disclosure of sensitive individual information contained within the underlying data. To this end, the most crucial step of PPDI is *Privacy-Preserving Record Linkage* (PPRL), which involves identifying and linking records about the same individual among multiple sources while avoiding privacy disclosure. The main elements of the PPDI and PPRL taxonomy and the various scenarios are detailed in Section 3. Section 4 presents a broad spectrum of related works and existing frameworks in different applications.

Building on these aspects, Section 5 defines the Recidivism Data Mart Project, the related privacy scenario, and specific requirements. Our main contribution to the RDM project was the de-

sign of a *Privacy-Preserving Data Integration* (PPDI) framework to be used as a basis to provide unified access to Italian legal-related sources. Among them, the most significant are: *Judicial Records* (indictments, verdicts, and other legal proceedings), *Prison Records* (sentence lengths and incarceration details) *External Penal Execution Records* (information on sentences served outside of prison, such as community service or parole). The distinct steps of the PPDI process, developed to satisfy the unique requirements of the RDM project resulting from the convergence of Italian legal-related data, European privacy regulation, and integration demands to assess recidivism, are described in Section 6. To fulfill these requirements and optimize the trade-off between privacy and utility of data, we focused on the realization of a Proof of Concept (PoC) for the crucial step of *Privacy-Preserving Record Linkage* (PPRL), which is discussed in detail in section 5. Additionally, the contributions and lessons learned from the study are described in Section 7, along with suggestions for future research directions and developments.

2. PRIVACY-PRESERVING DATA INTEGRATION

The Recidivism Data Mart Project aims to integrate criminal and court sources within the Italian Justice Domain to establish a unified Data Mart and assess the recidivism phenomena. In this section, we first outline the *Data Integration* (DI) process necessary to incorporate different autonomous sources; then, we present the privacy requirements concerning the processing of data about individuals in compliance with GDPR and introduce the concept of *Privacy-Preserving Data Integration*, discussing the DI process within a privacy context.

2.1 Data Integration Process

Data Integration (DI) is the process of consolidating data from a set of heterogeneous data sources into a single uniform dataset. To this end, the DI process in general involves three steps:

—**Schema Matching :** To produce a unified view of multiple data sources, it is necessary to develop an integrated conceptual schema of the different local schemas. Schema Matching resolves inconsistencies by finding the correspondences among the Local Sources and producing an integrated Global Schema.

—**Record Linkage :** Multiple data sources and multiple records within a single data source may describe the same real-world entity. Record Linkage (RL) (a.k.a. *Entity Resolution (ER)*) resolves inconsistencies at the tuple level by identifying and linking records that refer to the same real-world object.

—**Data Fusion:** Combines linked records from different sources into a single, consistent record by resolving conflicts in shared attribute values and creating a unique record for each individual.

2.2 GDPR

Whenever sensitive personal data about individuals are to be integrated, privacy and confidentiality implications have to be considered. Data protection in Europe is regulated by the *European General Data Protection Regulation (GDPR)*. Regarding PPDI, the GDPR leads toward the adoption of general IT security practice and specific techniques [2] to prevent internal parties involved in the PPDI process and external adversaries from the possibility of identifying an individual, called *Re-identification*. *Anonymization* is the process of removing any identifying information of an individual from the data in such a way that individuals become permanently unidentifiable. *Pseudonymization* [2] is the process of replacing identifying information with a *pseudonym* in such a way that additional information is needed to re-identify the individual. In addition, the additional information held separately can made available under controlled conditions for permitted re-identification of individual data subjects. For example under GDPR, if the controller becomes aware of a personal data breach it must identify the data subject and report the breach.
To apply the aforementioned techniques, the PPDI process is based on the classification of data content according to the concepts of *identifiability* and *privacy* established by the GDPR:

—*Personally Identifiable Information* (PII) denotes attributes that hold the potential to identify an individual. These include direct PII (e.g. identification number) and indirect PII or *Quasi-IDentifiers* (QID) that can identify a specific individual when combined (e.g. name, surname, date of birth, and address).

—*Sensitive Personal Information* (SPI) denotes confidential personal attributes to be protected from privacy disclosure (e.g. medical history or criminal records).

—*Non-Sensitive Data*: denotes attributes that contain neither identifying information nor information which deserves protection (e.g. metadata).

2.3 Privacy-Pseserving Data Integration Process

Privacy-Preserving Data Integration (PPDI) Process [4] is a branch of Data Science focused on providing a unified and accurate representation of personal information across multiple heterogeneous data sources while preventing the disclosure of individuals' privacy contained in the underlying data. In the following, we discuss how the GDPR requirements and the classification of data content is considered in the three steps of the PPDI process:

—**Schema Matching** Within a privacy context, local source schemas are generally available in plaintext. Therefore, traditional schema matching methods can be employed. However, the sets of PII and SPI are considered disjointed in a PPDI process and undergo distinct procedures. Thus, the Schema Matching phase typically entails the classification of the local schemas based on identifiability and privacy. This is precisely the situation in our project, as will be detailed further in Section 6.

—**Privacy-Preserving Record Linkage (PPRL)**
PPRL aims to develop techniques that enable the linkage of records without revealing sensitive or confidential information about the represented individuals. To this end, record linkage is based on Personally Identifiable Information (PII), which under the GDPR undergoes specific pseudonymization techniques to allow linkage while preventing re-identification. PPRL is the central focus of our project and will be discussed in section 6.4.

—**Data Fusion** is generally performed only on Sensitive Personal Information (SPI) accessible in plain format to allow further analysis. In a privacy context, SPI could be linked to external information containing identifiers. To reduce the possibility of re-identification, it is advisable to apply *Statistical Disclosure Control* methods [7], such as k-anonymity and differential privacy to the fused dataset. This aspect is outside the scope of the RDM project, as the resulting integrated data were analyzed internally within the Data Mart for recidivism analysis.

3. PRIVACY-PRESERVING SCENARIO

The effectiveness of the PPDI process, particularly the PPRL phase, is influenced by various aspects [16]. The three most

important are how many parties are involved in the process, the techniques employed for pseudonymization, and the adversary model assumed.

Number of parties The first aspect characterizing the PPDI scenario is the number of parties:

—*two-party protocols* is the most basic scenario, which only involves the participation of two database owners (DOs) in the process. Two-party protocols often have low communication costs, but complex PPRL and pseudonymization techniques are required to ensure that the two DOs cannot infer sensitive information from each other during the linkage process.

—*Multi-party protocols* involve multiple Data Owners (DOs) collaborating in the linkage process. These protocols use pseudonymization to link data from more than two sources and identify matching record sets across all parties. A common example is Secure Multiparty Computation (SMC) [8]. However, they introduce additional challenges in terms of scalability, linkage accuracy, and security.

—*three-party protocols*, require the separation of the roles in the PPRL process such that no single party can access both SPI and PII. In this approach, DOs encode the QID using specific pseudonymization techniques and send them to a *Third Party* or *Linkage Unit* (LU) that conducts the linkage.
This is precisely the approach employed in the PoC (see Section 5).

Pseudonymization Techniques
The second aspect is the pseudonymization techniques employed. Multiple pseudonymization techniques have been specifically studied in the literature [16].
The state-of-the-art technique is *Bloom Filter* (BF) and its variations [11]. Bloom filter is a bit vector data structure into which values are mapped using a set of hash functions. Initially, all bits are set to 0, and then each element in a set is hashed or mapped into using a set of independent hash functions where the bit position returned by each hash function is set to 1. One variation involves using *Attribute-level BF* (ABF), where each quasi-identifying attribute (e.g., first name, last name) is associated with a distinct BF. This approach permits precise similarity calculations per attribute, enabling detailed classification of record pairs based on varying attribute weights. In contrast, a single BF can be generated per record using *Record-level BF* (RBF) methods. *RBF*, constructs individual ABFs per attribute, samples bits based on attribute weights, concatenates them into a single BF, and applies random permutation for enhanced privacy during comparison [5]. Another example is *cryptographic long-term key* (CLK) [12], which hashes all attribute values and performs logical operations to condense them into a single BF per record. Many other variations and encoding methods exist to meet different requirements and scenarios. The *Tabulation Min-Hash* (TMH) encoding method was proposed by Smith [13] as an alternative to BF encoding to provide both improved similarity and privacy protection on small data sets. The challenge lies in choosing the best techniques because it must consider several aspects, such as the nature of the data, computational requirements, and the performance and protection required. Considerations on the comparison of selected techniques are provided in Subsection 6.4.

Adversary Model The third aspect regards adversaries. The goal of an adversary is to increase his information, which represents a violation of privacy rights under GDPR. The most severe one is to re-identify certain entities. The types of adversaries are classified according to the information they possess:

—*Insider Adversaries* have specific knowledge, capabilities, or permissions about the target. For example, an insider could be on the database owner's side (e.g. a malicious employee).

—*External Adversaries* not have direct access to relevant information. However, they may have access to pseudonymization techniques and to the pseudonymized datasets.

An adversary could or could not follow a specific protocol when trying to increase his information about the target datasets. The extremes of behavior considered are:

—*Honest but Curious* model assumes that parties follow the protocol while being curious to find about another party's data.

—*Malicious model* assumes that parties or adversaries can behave arbitrarily.

Based on these criteria adversaries can perform different types of attacks. The effectiveness of the attack depends on several parameters, including the publicly available information, the background knowledge of the adversary, the PPDI scenario and the possibility of collusion with other parties, and the pseudonymization technique employed. Different functions techniques have different privacy vulnerabilities and are susceptible to different adversary attacks [17]. For example, dictionary attacks are possible for hash-based pseudonymization functions, while frequency attacks are for attribute-level pseudonymization functions.

4. RELATED WORKS

The key aspects presented in Section 3 are referenced in a substantial body of literature in the context of PPDI and PPRL [10, 16, 17]. Regarding the number of parties involved, a thorough analysis conducted in [10] concludes that two-party and multi-party protocols, including those based on Secure Multiparty Computation (SMC) [8], are unfeasible for real-world decentralized applications. For example, in most European countries, medical data and criminal records are distributed among many different agencies, which do not allow an external Internet connection for their databases. These secure environments are not suited to protocols that require repeated access to external servers or several network interactions between database owners.
The most widely used and analyzed approach in the literature is based on a *Third Party* (TP) that conducts the linkage. As discussed in [3], in the traditional TP approach database owners (DO) send plaintext personal identifier information (PII) to the TP, which provides the linkage result. These cases are based on the concept of a *Trusted* Third Party (TTP). However, the GDPR leads toward the adoption of pseudonymization techniques to prevent both the calculation and the output of the calculation from permitting the possibility of identifying a specific individual. In more recent and decentralized PPRL protocols, each DO encodes the PII using pseudonymization techniques, ensuring that no plaintext identifiers are exposed. The TP then performs linkage using pseudonyms, preventing any sensitive information from being disclosed. The distinction between trusted and untrusted third party is not always clear-cut in concrete application approaches, and intermediate solutions exist. In Chapter 13 of [3], several real-world applications of linking sensitive databases developed by different countries are described. For instance, Brazil increasingly adopted PPRL, including scenarios involving untrusted third parties that employ advanced

encoding techniques. On the other hand, data protection legislation in Australia has allowed the extensive use of trusted third-party methods based on plain-text QID and probabilistic linkage methods. However, for cases where QID cannot be safely disclosed (such as names and other personal data from the Children's Court), pseudonym-based techniques are implemented.
The main European data integration projects under GDPR are related to the Health Domain.
The *European Patient Identity Management* (EUPID) [1] prevents duplication of patients and allows linkage of data related to the same patient for secondary use, avoiding the creation of a universal patient ID. In addition, EUPID enables to inform patients about relevant research results and data breaches. The *Secure Privacy-preserving Identity management in Distributed Environments for Research* (SPIDER) pseudonymization tool allows linkage of data related to the same patient while avoiding re-identification of the patient's identity and preventing unencrypted data from leaving the local storage. SPIDER is feasible in a distributed computing environment. Every European citizen has a European Health Insurance Card - EHIC, which can be used as direct PII, facilitating matching between individuals. For this reason, the Health Domain does not necessitate an advanced PPRL process; rather, it necessitates general IT security practices. In Europe, Germany primarily faced PPDI challenges related to the Justice Domain and explored both trusted and untrusted models for data linkage [3]. The majority of the scenario, however, involved compliance with national privacy policies in place of GDPR. Italy, on the other hand, has not adequately advanced PPDI projects. However, some articles have considered complementary aspects to our research project within the Italian justice Domain. [9] tackles the extraction and management of named entities within Italian civil court judgments using *Natural Language Processing* (NLP) techniques and annotation pipelines. Their focus is on optimizing results and overcoming challenges related to the scarcity of annotated data. In both works, each organisation annotates and extracts metadata from its dataset. Our research addresses the subsequent step, namely how to perform privacy-preserving data integration of the extracted metadata datasets.

5. RECIDIVISM DATA MART PROJECT

The work discussed in this paper is part of the "Recidivism Data Mart and Criminal Data Warehouse" project. Our objective was the design of a PPDI framework for the Italian Justice Domain. The framework is illustrated in Figure 1, and will be discussed in the following.

5.1 Project Scenario and Requirements

In this section, we consider the privacy aspects discussed in 3 and outline the scenario and requirements specific to the RDM project. The goal of the RDM project is to integrate Italian criminal and court sources to assess recidivism phenomena. To maximize the trade-off between privacy protection and the utility of the data for the recidivism analysis is necessary the de-duplication of criminal datasets in compliance with GDPR. Section 6 details how the PPRL process was actually implemented (see Figure 2).
The first requirement of the project consists of its decentralized nature. Crime records are distributed among many different parties that do not allow external internet connections for their databases. The adversary model considered is both internal (e.g. the sources involved in the process) and external adversaries with honest-but-curious behaviour. As described in Section 2, to carry out the PPDI process efficiently in a specific scenario, it is necessary to classify a priori PII and SPI. The first problem to consider is the absence of direct PII among the RDM sources to be linked; consequently, linkage techniques must rely on the use of Quasi-Identifier (QID) attributes. Therefore, in order to perform Record Linkage in conformity with GDPR, it is important to apply advanced data pseudonymization techniques for the QID attributes and tolerant matching approaches.
The architecture we adopted to meet the aforementioned requirements is represented in Figure 1. The concept that served as the starting point to design the PPDI framework architecture is the Third-Party approach, which represents a reference in the literature in the context of decentralized organizations, where legal requirements limit the number of applicable approaches (explained in detail in Section 4). As shown in Fig. 1, the TTP will serve as the PPDI Domain to provide the Consumer Domain with a unified and privacy-preserving representation of the different autonomous data sources within the Source Domain. The basic communication steps between the parties can be summarized as:
- exchanging of functions and parameter values,
- sending of the (masked) data of the databases,
- sharing of the aggregated results.

This architecture principle was coupled with the *Linkage Unit* approach, discussed in [10] and particularly well-suited for conducting PPRL in decentralized organizations with multiple data sources. As highlighted in [3], another key advantage of the *Linkage Unit* approach is its ability to implement the *separation principle*. This principle divides the responsibilities involved in the PPRL process to ensure no single party can access to both QID and SPI.
- *Decision Unit*: defines the set of QID for record linkage;
- *Linkage Unit*: manages the QID required for record linkage without accessing SPI;
- *Data Fusion Unit* manages only the SPI to create a unique record for each real-world entity.

In the original architecture proposed in [10], the Linkage Unit (LU) is a simple Third Party. However, our architecture introduces a significant change by considering the LU as a Trusted Third Party (TTP). This change is driven by the project's requirement to comply with GDPR, which include provisions for re-identifying data subjects in case of a personal data breach. Our architecture provides pseudonymized data to be transmitted from local sources to the TTP in order to comply with GDPR regulations. Nevertheless, in cases controlled by GDPR when the re-identification is mandatory, the TTP can request plain-text data from the local source, enabling the permitted re-identification procedure. It's important to note that all communication channels are encrypted to respect IT data security practices. The analysis of related works (see Section 4) did not find similar cases where a Linkage Unit needs to be Trusted to comply with GDPR re-identification requirements. Our PoC, described in Section 6, demonstrates that this approach is feasible.

6. PROOF OF CONCEPT

In this section, we provide a detailed discussion of the Proof of Concept (PoC) and of the PPDI process to meet all the specific requirements of the RDM project described in Section 5.
A major limitation to PPRL research projects based on concrete application cases is the inability of organizations to share real data as it is protected under the GDPR. To this end, Poc is very significant as it is based on real sources from the Italian Justice Domain. Subsection 6.1 provides an overview of the three most significant legal data sources employed. However, one limitation of the RDM project is that organizations were only allowed to share the original local schemas of the sources, from which the Schema Matching

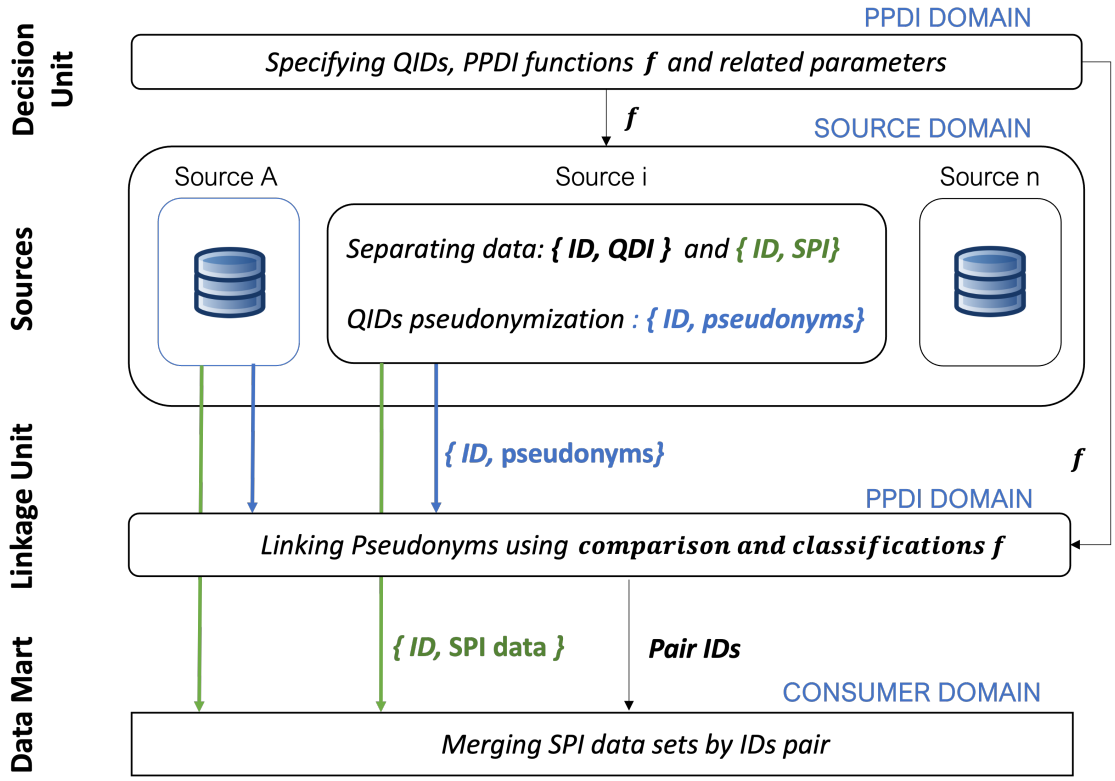

Fig. 1. Schema of the PPDI Architecture

phase, described in Subsection 6.2, was carried out. Nevertheless, a synthetic dataset had to be created to realize the Proof of Concept for the PPRL process. In subsection 6.3, we describe the generation of the Italian Justice synthetic datasets, based on the source schemas and respecting the distribution and characteristics of real data. The focus of the PoC was the PPRL process implementation, which is extensively discussed and exemplified in Subsection 6.4.

6.1 Description of the Sources

The data sources incorporated into the RDM project can be broadly categorized into two groups: those internal to the justice domain (divided by the responsible Department) and those external to the domain (divided by the responsible Public Administration). The sources were numerous and the real schemas were complex, consisting of hundreds of tables and attributes. In this section, we briefly describe only three significant sources of the project:

—*Judicial Records from the Department of Justice Affairs* (S_1): This source includes information on charges (Indictments) and legal outcomes (Verdicts).

—*Prison Records Information System* (S_2): Records the duration of sentences served in prison.

—*External Penal Execution Information System* (S_3): Tracks sentences served outside prison, such as community service.

6.2 Schema Matching and QID Specification

The schema matching step involves selecting, from various local conceptual schemas, the tables and attributes relevant for linkage and analysis. The selected schemas are then compared to identify correspondences between local attributes, resulting in the construction of a global integrated schema. These correspondences define the new entities and relationships of the global conceptual schema. Schema matching is typically a complex, time-consuming, and subjective task, often performed manually by domain experts.

As discussed in Section 5.1, in privacy-preserving contexts it is also necessary to classify Personally Identifiable Information (PII) and Sensitive Personal Information (SPI) in advance. In the case of the RDM sources, the absence of direct PII required the use of Quasi-Identifier (QID) attributes. For effective linkage, a QID must be present in all local sources and its combination (or a subset thereof) must uniquely identify the entities.

In the PoC, due to the possibility of accessing the real local sources schemas, the process to categorize and establish correspondences between the PPI and SPI of the local sources was performed manually. As a result, a Global Schema across all Italian legal sources was produced, along with the Data Transformation Functions to allow the schema alignment between local and Global schema. In addition, a subset of QID (common to all sources) was selected to carry out the PPRL phase.

To exemplify this process we consider some selected QID belonging to RDM sources S_1, S_2, and S_3, shown in the following table, where the first column contains the attributes A_i of the *QID-Global Schema*, and the corresponding element to (A_i, S_j) represents the set of local attributes from source S_j that are mapped to A_i.

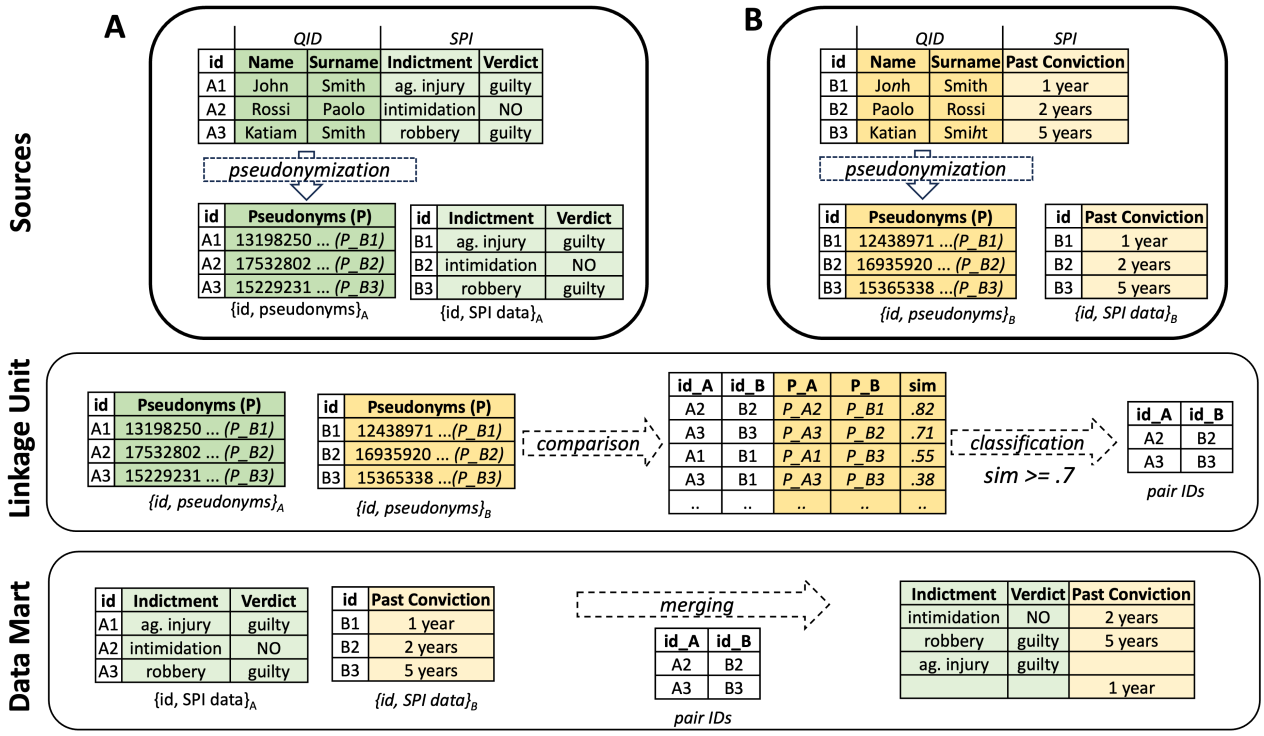

Fig. 2. Example related to the PPDI Architecture

QID-GS	S1	S2	S3
Name	`Name`	`Full Name`	`Name`
Surname	`Surname`	`Full Name`	`Surname`
Gender	`Gender`	`Gender`	`Gender`
DOB	`Y, M, D`	`DateOfBirth`	`DateOfBirth`
POB	`BirthPlace`	`CodiceBelfiore`	`BirthPlace`

The process can be summarized as:

(1) Identify correspondences between local sources. For example, between (*Name*, S_1) and (*Full Name*, S_2) and between (*Surname*, S_1) and (*Full Name*, S_2).

(2) Select a set of QID common to all sources, called the *QID-Global Schema* (QID-CS):
QID-GS = {*Name*, *Surname*, *Gender*, *DOB*, *POB* }.

(3) Map the local QID to the *QID-Global Schema* and define the transformation functions to return a common format. For example, split (*Full Name*, S_2) into *Name* and *Surname*.

Nevertheless, correspondences and transformation functions are not as straightforward in real projects. For example, the *POB* Global attribute represents the place of birth of the criminal. This information is associated with the so-called *"Codice Belfiore"* of source S_2, with the following meanings:
- For citizens born in Italy: a four-character alphanumeric code that uniquely identifies the municipality or subdivision (*frazione*) of birth within Italy.
- For citizens born abroad: the foreign country where the individual was born.
In the other two sources *BirthPlace* represents a string containing information about *location* and *country*. However, the documentation specifies that for foreign nationals the *location* field is optional and may be absent. Therefore, to obtain consistent information across all sources, *"Belfiore Code"* is chosen as the common format. Appropriate transformation functions are used to derive this code from the strings of sources S_1 and S_3. After applying this functions all QID will have the same name (as shown in the first column of the above table) and the same format.
Furthermore, QID are not stable over time and data values may be subject to recording errors and missing values.
For this reason, the first step of the PPRL process (see 6.4), which must be performed locally and independently within each source, is the **Pre-processing of raw data**. This involves applying the corresponding transformation functions to return QID into a common format and using attribute-specific functions to normalize error-prone QID and to prepare sensitive data for analysis.

6.3 The Synthetic Dataset "AnagraficaGiustizia"

In the PoC, synthetic datasets were used, which include the QID attributes from the QID-GlobalSchema and are designed to be as realistic as possible. The process followed is as follows:

—**Dataset Creation**: A dataset simulating the demographic records of prisoners in Italy was created, based on statistics from the Italian Ministry of Justice and ISTAT. Attributes like Name and Surname were generated using the Faker tool, while Sex, Place of Birth, and Date of Birth were randomly generated according to ISTAT's distribution statistics in Italian prisons. This first dataset *DA* contains 2,000 records.

—**Dataset Corruption**: The second and third datasets, DB and DC respectively, were obtained from DA through specific corruption operations to simulate real-world "dirty" data. For this purpose, we used the GeCo tool [14], which allows various types of changes to be applied. These corruptions simulate data entry processes that can lead to manual typing errors, scanning errors, and OCR inaccuracies.

—**Overlap Adjustment**: We produced corrupted datasets (DB and DC) in which every record was required to be a duplicate of a record in DA. Further processing of these two datasets was necessary to achieve arbitrary degrees of overlap between DA, DB, and DC. As a result, there may be records in both DA, DB, and DC that do not have any duplicates at all. This approach allows us to obtain a more realistic scenario.

6.4 Privacy-Preserving Record Linkage process

The PPRL process depends on various factors, discussed in Section 3, but essentially follows the same steps as Record Linkage performed in traditional Data Integration 2.1, applied to pseudonymized *QID* values. In the POC, the focus was on the techniques employed for pseudonymization and the corresponding methods for comparing and linking the obtained pseudonyms, described in Subsection *6.4.1*.
As discussed in Section 5.1, our *Trusted Third Party* architecture (Figure 1) leverages the *Linkage Unit* to implement the *separation principle* and efficiently link records without compromising privacy. Our project, shown in Figure 2, envisions that the different steps of the PPRL process are performed by different parties to ensure that no single internal party has access to the totality of background information nor can access both QID and SPI.

—The *Decision Unit* is represented by the researchers who carried out the matching phase, defining the set of QID required for record linkage, the related transformation/normalization functions and the pseudonymization techniques to be used in the PPRL process.

—The *Linkage Unit* is represented by the PoC which implements the comparison and linkage of the pseudonymized *QID* values.

—The *Data Fusion Unit* is represented by the resulting Recidivism Data Mart which contains the aggregated dataset with the SPI values for each real-world entity.

Therefore, the steps of the PPRL process and the respective party in charge are as follows:

—**Pre-processing** (performed by each source)
Using the transformation functions, error-prone QID is transformed into a unique and comparable format, and SPI is transformed into a format that is useful for the analysis.

—**Pseudonymization** (performed by each source)
Using pseudonymization methods, common-format QID are transformed into *pseudonyms* to allow linkage while preventing re-identification. In addition, specific transformations related to the chosen technique can be employed. For example, the following subsection will describe in detail two pseudonymization methods that require the concatenation of QID.

The local sources then send record_ID and Pseudonym for each record to the Linkage Unit; to enable the subsequent PPRL steps. Before the comparison step it is possible to perform the **Blocking** step. Blocking is needed only to face scalability issues, as it reduces the number of comparisons to be conducted producing candidate pseudonym pairs of the pseudonyms that are likely to match (this was not the case in the PoC). In the context of PPRL, blocking can either be conducted locally within the sources using plain-text QID values or by the Linkage Unit using pseudonyms.

—**Linking** (performed by the *Linkage Unit*)

—**Comparison**: using approximate similarity functions to compare pseudonym pairs. The choice of similarity function depends on the pseudonymization method used. The most commonly used functions are classic ones, namely *Jaccard similarity* and *Dice distance*.

—**Classification**: using a decision model based on the results of the comparison to classify candidate pseudonym pairs into matches or non-matches. A classic model employed for this purpose is the threshold-based model.

The output of the PPRL process is the record_ID pairs classified as matches, (i.e. referring to the same real-world entity). To produce an integrated representation of the phenomenon of recidivism, it is necessary to develop an aggregated database, which allows a user to pose a query and receive a single, unified answer. To this end, the local sources then send record_ID and SPI data for each record to the Fusion Unit and the Linkage Unit sends the pairs of matching record_ID. The next stage is the aggregation of SPI for each group of matching record_ID (representing the same real-world entity) to produce a global record. In the RDM project, the resulting integrated SPI were simply concatenated and stored in the so-called Recidivism Data Mart for internal use in the recidivism analysis. It is worth noting that in compliance with the GDPR, the RDM was anonymized (also the record_ID was removed) and completely separated from the Metadata Table. The Metadata Table is the storage of the record_ID and Pseudonym to allow the possibility of controlled re-identification. In this way the LU can request plain-text data from the local source about the specific Pseudonym, enabling the permitted re-identification procedure.

6.4.1 Pseudonimization. The pseudonymization of quasi-identifying attributes (QID) is performed locally by each data source, adhering to the guidelines set by the Decision Unit. In the Proof of Concept, we chose to use well-established pseudonymization techniques, that have been thoroughly evaluated in the literature [16, 11, 12, 13] to focus on their applicability in our specific context. Considering the techniques described in Subsection 3 we prevented the use of *Attribute-level BF* (ABF). This approach has a significant drawback in real world scenario: common attribute values result in identical bit patterns, rendering the BFs susceptible to frequency-based attacks.
Frequency attack assumes the adversary has some knowledge about the type of values encoded in a database. A frequency attack is based on the frequency distribution of a set of encoded values which is matched with the distribution of known unmasked values in order to infer the original values of the masked values.
Therefore, we decide to use *Record-level BF* (RBF) variations. One of the techniques employed is the *Cryptographic Long-term Key* (CLK), which allows for the assignment of different weights to attributes [12]. In this method, all QID (such as first name, last name, sex, etc.) are mapped to a single Bloom filter using distinct hash functions. By varying the number of hash functions applied to each QID, we can reflect the assumed discriminatory power of identifiers for potential record pairs. For instance, when calculating similarities between records, we may assign greater importance to the last name compared to the first name, as last names often carry more weight in identifying individuals, particularly in systems that involve personal identification or demographic records.

Our aim was to gain insights into their implementation within the project context, considering the primary characteristics of the dataset. Through our tests, we observed a key characteristic of the *CLK* method: it proves to be highly effective when dealing with high-quality data where attributes are accurate and consistent across various datasets. However, its performance can significantly decline in the presence of data errors, such as swapped first and last names or missing values. In such scenarios, it may be more suitable to consider pseudonymization techniques that treat all QID attributes as a single entity: the values of these attributes are concatenated into a string that is encoded to generate a single pseudonym.
As an encoding technique, we employed the *tabulation-based min-hash* (TMH) method [13]. This method, proposed as an alternative to Bloom filter encoding, enhances similarity detection for long-valued attributes while ensuring privacy protection, particularly for smaller datasets. For brevity, implementation details and results are reported in the technical report [15].

Evaluation Insights
Our evaluation included both the synthetic dataset described in Section 6.3 and the North Carolina Voter Registration (NCVR) dataset, one of the most commonly used benchmarks in the literature for assessing both traditional RL and PPRL techniques. Overall, the pseudonymization techniques we employed yielded good performance on both datasets. The most relevant results were obtained on the NCVR dataset, where the CLK method achieved high accuracy, with precision and recall values typically ranging between 0.95 and 0.99, depending on the parameter settings. These parameters include, for example, the number of hash functions, the length of the Bloom filter, and the selection of quasi-identifiers. While the primary goal of our PoC was not to conduct a comprehensive evaluation of PPRL methods, the results we obtained are in line with those reported in the literature [3], and confirm the effectiveness of the techniques in practical scenarios.

7. CONCLUSION AND FUTURE WORKS

In this paper, we presented a framework for PPDI aimed at supporting recidivism analysis within the Italian Justice Domain. By integrating criminal and court records from autonomous sources, the framework enables a unified and comprehensive view of individuals' criminal histories. A key aspect is the use of PPRL, which allows linking related records without exposing sensitive information, thus ensuring compliance with privacy regulations such as the GDPR. We developed a PoC to demonstrate the feasibility of the approach, focusing particularly on the pseudonymization techniques required for privacy protection during data integration.
As future work, we plan to explore Privacy-Preserving Data Publishing (PPDP) techniques [6], such as differential privacy, to further reduce the risk of re-identification when data is shared externally. This would allow external entities to analyze anonymized data using advanced data-centric AI techniques.
More broadly, we believe that addressing both data integration and data publishing in a unified framework is essential to safeguard privacy. Our ongoing efforts aim to extend the framework to handle diverse real-world scenarios, including the development of tools to classify data based on identifiability, support schema matching, and select appropriate privacy-preserving strategies.
A key challenge remains the lack of empirical metrics to assess the trade-off between data utility and privacy. Tackling this issue will be central to our future research.

8. REFERENCES

[1] Eupid - european patient identity management. `https://eupid.eu`. Accessed: Aug. 01, 2024.

[2] Luca Bolognini and Camilla Bistolfi. Pseudonymization and impacts of big (personal/anonymous) data processing in the transition from the directive 95/46/ec to the new EU general data protection regulation. *Comput. Law Secur. Rev.*, 33(2):171–181, 2017.

[3] Peter Christen, Thilina Ranbaduge, and Rainer Schnell. *Linking Sensitive Data - Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer, 2020.

[4] Chris Clifton, Murat Kantarcioglu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed K. Elmagarmid, and Dan Suciu. Privacy-preserving data integration and sharing. In *DMKD*, pages 19–26. ACM, 2004.

[5] Elizabeth Ashley Durham, Murat Kantarcioglu, Yuan Xue, Csaba Tóth, Mehmet Kuzu, and Bradley A. Malin. Composite bloom filters for secure record linkage. *IEEE Trans. Knowl. Data Eng.*, 26(12):2956–2968, 2014.

[6] Benjamin C. M. Fung et al. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.

[7] J M Gouweleeuw, Peter Kooiman, and PP De Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of official Statistics*, 14(4):463, 1998.

[8] Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *J. Priv. Confidentiality*, 1(1), 2009.

[9] Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. Named entity recognition and linking for entity extraction from italian civil judgements. In *Proceedings of AIxIA 2023*. Springer, 2023.

[10] R. Schnell. Privacy-preserving record linkage. In K. Harron, H. Goldstein, and C. Dibben, editors, *Methodological Developments in Data Linkage*, pages 201–225. John Wiley & Sons, UK, December 2015.

[11] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics Decis. Mak.*, 9:41, 2009.

[12] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. A novel error-tolerant anonymous linking code, 2011.

[13] Duncan Smith. Secure pseudonymisation for privacy-preserving probabilistic record linkage. *J. Inf. Secur. Appl.*, 34:271–279, 2017.

[14] Khoi-Nguyen Tran, Dinusha Vatsalan, and Peter Christen. Geco: an online personal data generator and corruptor. In *Proc. of CIKM 2013*, pages 2473–2476. ACM, 2013.

[15] Lisa Trigiante, Domenico Beneventano, and Sonia Bergamaschi. Privacy-preserving data integration for digital justice. Technical Report TR-2025-01, Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, 2025. `https://dbgroup.ing.unimore.it/TecRep/trigiante2025_1.pdf`.

[16] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Inf. Syst.*, 38(6):946–969, 2013.

[17] Anushka Vidanage, Thilina Ranbaduge, Peter Christen, and Rainer Schnell. Taxonomy of attacks on privacy-preserving record linkage. *J. Priv. Confidentiality*, 12(1), 2022.