[Vision Paper] Privacy-Preserving Data Integration

Lisa Trigiante* , Domenico Beneventano and Sonia Bergamaschi
DBGroup - Department of Engineering Enzo Ferrari - University of Modena and Reggio Emilia
Modena, Italy
{*name.surname*}@unimore.it

***Abstract*—The digital transformation of different processes and the resulting availability of vast amounts of data describing people and their behaviors offer significant promise to advance multiple research areas and enhance both the public and private sectors. Exploiting the full potential of this vision requires a unified representation of different autonomous data sources to facilitate detailed data analysis capacity. Collecting and processing sensitive data about individuals leads to consideration of privacy requirements and confidentiality concerns. This vision paper provides a concise overview of the research field concerning Privacy-Preserving Data Integration (PPDI), the associated challenges, opportunities, and unexplored aspects, with the primary aim of designing a novel and comprehensive PPDI framework based on a Trusted Third-Party microservices architecture.**

***Index Terms*—PPDI, PPRL, Pseudonymization, Big Data**

The digitization of legal, administrative, and healthcare processes, among many others, has generated vast amounts of data describing people and their behavior. The resulting person-related Big Data presents substantial intrinsic worth and holds considerable potential to feed multiple research areas with the aim of enhancing the human condition. Achieving this vision requires an efficient *Data Integration* (DI) process to create a unified view of different data sources and enable in-depth and extensive analysis that is infeasible through any individual source. To this end, the DI process involves three steps:

- **Schema Alignment** resolves inconsistencies at the schema level by finding the semantic correspondences among the schema of the Local Sources and producing an integrated Global Schema.
- **Record Linkage (RL)** resolves inconsistencies at the tuple level by identifying records about the same individuals from different sources.
- **Data Fusion** resolves inconsistencies at the value level by creating a unique record for each individual.

However, privacy and ethical implications pose a serious challenge to the integration of data about individuals.

Privacy-Preserving Data Integration (PPDI) [1] is a branch of Data Science aimed at providing a unified and accurate representation of personal information across multiple heterogeneous data sources while preventing privacy disclosure of individuals represented in the underlying data.

This vision paper aims to provide a concise overview of the PPDI process and to delineate our current and future works within this research field, grounded in the existing body of literature. Specifically, Section I discusses the main challenges and opportunities arising at the convergence of the Big Data Integration process and privacy requirements. In Section II, we present our approach to the design of an innovative and comprehensive PPDI framework. Furthermore, we illustrate the methodology devised to facilitate each phase of the PPDI process and to address a range of related issues and uncharted aspects within the literature. Finally, in Section III, we conclude with the expected contributions of this research and provide insights for future directions and developments.

*ICT PhD Student founded by MIUR under D.M.351 with the Emilia Romagna region as partner.

I. PPDI Challenges and Opportunities

As Data Science meets the privacy context, domain-specific challenges and opportunities materialize in various research fields. The exploitation of Data Mining and Artificial Intelligence (AI) techniques presents the opportunity to advance public and private systems toward innovative Data-Driven approaches that support the emerging 5P (Predictive, Preventive, Personalized, Participatory, and Precision) paradigm [2]. However, per the Data-Centric principle, the quality and quantity of data used to train AI models are critical factors in determining their analysis capacity and accuracy performance.

A. BIG DATA ISSUES

Big Data [3] presents a severe challenge in terms of volume, velocity, variety and veracity (4V of Big Data) and often exhibits intrinsic issues and a sparse, scarce, and unbalanced nature. Numerous approaches are available within the Data Integration literature to address different facets of data quality [4]. However, the incorporation of privacy requirements pose additional challenges and necessitates the modification of traditional process through the adaptation of pre-existing approaches and methodologies and the development of innovative privacy-preserving techniques. These specifications impose strict demands on the data resulting from the PPDI process with respect to scalability, completeness, balance among target classes, consistency, and regularity over time.

B. GDPR REQUIREMENT

The European *General Data Protection Regulation* (GDPR) bases the classification of data content on the concepts of identifiability and privacy:

Personally Identifiable Information (PII) refers to attributes that identify an individual. These include direct PPI (e.g. identification number) and indirect PPI or *Quasi-IDentifiers* (QID) that have the ability to identify a specific individual when combined (e.g. name and address).

Sensitive Personal Information (SPI) denotes confidential personal attributes to protect from privacy disclosure, e.g. medical history or criminal records.

The GDPR leads toward the adoption of specific techniques [5] to prevent internal parties involved in the PPDI process and external adversaries from the possibility of identifying a specific individual, called *Re-identification*. For this reason, the sets of PII and SPI are typically considered disjointed in a PPDI process. SPI constitutes the outcome of the process and is accessible in plain format to allow further analysis. Conversely, PII undergoes specific pseudonymization techniques to facilitate record linkage while preventing re-identification. *Anonymization* is the process of removing PII from the data. *Pseudonymization* [6] replaces PII with a *pseudonym* (or encrypted code), to allow further processing.

C. PRIVACY AND USABILITY TRADE-OFF

In real-world scenarios, with any information disclosure there is always some privacy loss, and with any pseudonymization technique there is always some information loss. An important issue of *Statistical Disclosure Control* (SDC) is to ensure the optimal trade-off between measures to maximize the utility of data to be disclosed (which is equivalent to minimizing information loss due to the application of SDC methods) and to maximize privacy protection (i.e. to minimize the risk of re-identification).

There is an extensive literature base on information loss metrics and disclosure control methodologies [7]. In comparison, the evaluation of privacy is a big impediment as it represents the resistance to re-identification attacks and depends on aspects that are complex to quantify, such as the nature of the data involved and the publicly available information, as well as the different behaviors and knowledge bases of the adversaries [8].

For these reasons, the determination of a set of standard measures for the empirical evaluation of the trade-off between privacy and usability of data is still a developing area of the literature that necessitates careful consideration.

D. TEMPORAL ASPECT

A predominant assumption within the literature pertains to the static nature of datasets. Nevertheless, in a real-world context, this is not reasonable due to the dynamic nature of data, which can exhibit varying degrees of granularity. At the source level, there is the possibility of sources becoming unavailable or the inclusion of new information from additional sources. At the record level, the constant evolution of public and private processes and data collection methodologies warrants consideration as source schema may change and records may undergo modifications. At the attribute level, it is essential to recognize the inherent variability of data about individuals, e.g. surnames or addresses may undergo alterations over time.

II. COMPREHENSIVE PPDI FRAMEWORK

In this section, we present a methodology devised to support the creation of a novel and comprehensive PPDI framework, supported by the existing literature.

Our research in the field of PPDI has encompassed concrete application projects, such as the design and development of a Proof of Concept for the *Recidivism Data Mart and Criminal Data Warehouse* project, establishing a PPDI process across Italian legal data sources to assess the recidivism phenomena [9]. It is worth noticing that we intended from the design stage to accommodate different application scenarios and not tailor solutions to the justice domain. This approach sheds light on complex theoretical and practical issues (outlined in I and referenced below) when designing a PPDI to be effective in different domains. For instance, collaborative projects with the Health Departments of the Emilia Romagna region have underscored the challenges inherent in privacy-preserving processing of specific health-related data [10]. A more in-depth discussion of this study was addressed in [11].

In this paper, our objective is not to deeply delve into the discussion of state-of-the-art privacy-preserving techniques nor into specific issues and related research, for which we refer readers to the existing literature [12], or to delve into technical specifics. Instead, our aim is to offer a holistic approach to this research field. The intention is to discern uncharted domains or gaps that, to the best of our knowledge, have not received extensive coverage in the existing literature. Furthermore, we aim to shed light on insights for addressing these unexplored areas and present an overview of the current state of our research efforts for the design of a novel PPDI framework, including architectural approach and specific methodologies formulated for each stage.

The concept that served as the starting point [9] is a Trusted Third-Party (TTP) architecture, which represents a reference in the literature [13] in the context of decentralized organizations, where legal requirements limit the number of applicable approaches. To illustrate this point, Secure Multiparty Computation (SMC) [14] protocols, a potential alternative, are often unfeasible for real-world applications due to their impracticality for large datasets (see I-A), characterized by lengthy computing times and the need for several network interactions between internal parties. The idea behind the PPDI framework is an incremental extension of the *MOMIS (Mediator envirOnment for Multiple Information Sources)* [15] Data Integration system toward a TTP-based microservice architecture, including specific software modules to realize PPDI in compliance with the GDPR. 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 framework will provide different microservices designed to fulfill privacy requirements and specific functionalities, described above, for each step of the PPDI process.

A. Schema Alignment

Schema Alignment is the process of finding the correspondences between the different schema of Local sources and producing a unique integrated Global Schema. It is an extremely difficult, time-consuming, subjective, and intelligent process. Within the privacy context, this step is often taken

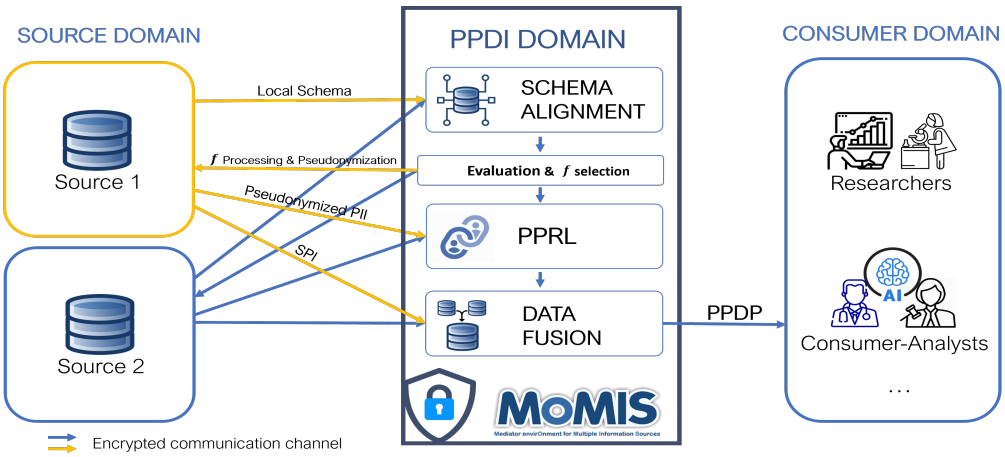

Fig. 1. Schema of the PPDI Architecture.

for granted as accessing schema or metadata in plain format does not pose significant privacy risks thus traditional schema-matching approaches can be employed, such as Linguistic, Instance-based, Structure-based, Constraint-based, Rule-based, Hybrid-matching. Nonetheless, these techniques consider local schema as a whole, whereas the PPDI process subjects PII and SPI to distinct processing procedures(see I-B). In the context of the PoC, as is common practice in the literature, a subset of QID common to all sources was established a priori by the human in the loop for use in the PPRL phase (detailed in II-B). However, this process is only feasible for low-dimensionality schemes and not in the context of Big Data (see I-A). Classifying data according to identifiability in a real-world scenario is a complex task because SPI and QID can overlap and may highly impact the overall trade-off between privacy and data usability (see I-C). The combination of attributes identifying an individual may vary from person to person depending on the rarity of attribute values.

In light of the aforementioned considerations, it is convenient to examine schema alignment techniques that leverage distinctions among schema elements to optimize the overall PPDI process and jointly provide the potential for assessing privacy preservation from this stage. It is worth noting that this distinction regarding attribute types has not been considered, to the best of our knowledge, in the existing literature on schema alignment techniques. Starting from our Data Integration system, MOMIS [15], we are investigating how to automatically distinguish QID and SPI (see I-B). Additionally, based on this, we aim to obtain hints to assess privacy and design the most appropriate PPDI techniques to be applied. Data instability over time (see I-D) leads to the design of a data integration process able to guarantee the Global Schema is modifiable and expandable over time.

B. Privacy-Preserving Record Linkage (PPRL)

PPRL is the process of identifying and linking records about the same real-world entities. It is the crucial step to achieve the best trade-off between privacy and usability as it strongly impacts the consistency and completeness of data resulting from the PPDI process. PPRL can be viewed as a classification problem that labels pairs of records across different sources like a match (i.e. two records refer to the same individual) or a non-match, However, the process to achieve high-quality results in a privacy-preserving setting, illustrated in Figure 2, comprises different steps:

1) *Pre-processing and Pseudonymization*
   Linkage of data about individuals is commonly based on QID since direct PII are often not present (and more vulnerable to re-identification attacks). However, QID is neither unique nor stable over time and may be subject to recording errors and missing values I-A. For this reason, pre-processing of QID toward a standard format is highly recommended [13]. In addition, compliance with the GDPR (see I-B) requires to prohibit non-pseudonymized QID from leaving the local storage. Therefore specific encryption techniques must be employed at source side to transform QID into pseudonyms.
2) *Linkage of pseudonyms*
   Subsequently, each local source transmits for each record the respective pseudonym (and SPI) to the TTP that performs the actual linkage in line with the traditional process employed in a non-privacy context. *Blocking* is an initial option step to address the challenge of computational and operational scalability (see I-A). Blocking techniques reduce the number of comparisons that need to be conducted by aggregating pseudonyms that exhibit a likelihood of correspondence into blocks and thereby generating candidate pseudonym pairs. Subsequently, the *Comparison* step evaluates candidate pseudonym pairs in detail using specific comparison functions and similarity measures. *Classification* of candidate pseudonym pairs into a match or not match is then performed using a decision model based on the results of the comparison. When local data sources are deduplicated, *Post-processing* methods [16] can be employed to refine linkage results to only one-to-one.

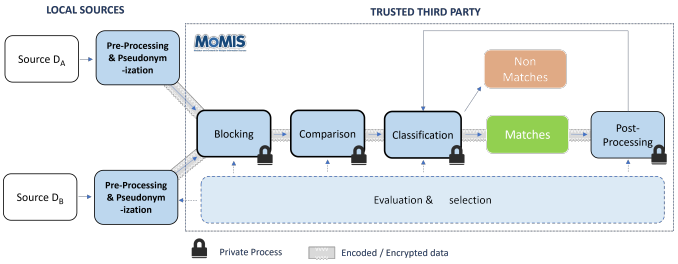

Fig. 2. Schema of the PPRL Process.

The usage of pseudonyms poses a significant challenge to the trade-off between privacy and performance (see I-C), as encrypted QID must ensure privacy preservation while allowing classification with minimal bias in comparison to plain-text record pairs. For example, a privacy technique for the encryption of a subject's DNA sequence that provides good protection is Fully Homomorphic Encryption (FHE) [17]. However, FHE has poor performance and a massive overhead in computational and memory costs.

From a broader perspective, evaluating the whole PPRL process is a complex issue because each step is dependent on and connected to the others. In order to allow the comparison and classification of pseudonyms is necessary that all local sources share a group of QID and that the Pre-processing and Pseudonymization phase is conducted exactly in the same way and order. To the best of our knowledge, there is no PPRL framework that considers sources with different QIDs, but only approaches that try to deal with missing values [18]. On the other hand, a lot of specialized techniques have been studied in the literature to cover the pseudonymization phase [19], and traditional approaches can be exploited to carry out the consecutive steps of the PPRL process.

For these reasons, we contend that the challenge lies in the selection of the best techniques to employ in each step as it must take into account different aspects, such as the nature of data, the computational requirements, the performance in terms of scalability and linkage quality and the protection achieved. However, this evaluation necessitates the actual record values (or ground truth) that are unlikely to be accessible as it would reveal private or confidential information.

In our perspective, the spectrum of interdependencies within PPRL and, from a more general viewpoint, PPDI underscores the imperative for a comprehensive framework that holistically addresses the entirety of the process. To achieve this goal, our project will exploit semantic knowledge extraction and schema and metadata annotation (see II-A) to infer the nature of data and delegate the TTP to select the set of QID alongside attribute-specific standardization and pseudonymization functions and parameters. This centralized vision facilitates the TTP to provide sources with a microservice implementing the first step of PPRL. This reduces IT and computational requirements at the source side and enhances the privacy of the process by decreasing the number of interactions and the background knowledge of internal adversaries. A comprehensive approach also empowers the TTP to select the best techniques to perform the related actual linkage, considering various factors that may guide the selection of advanced methods to address specific challenges of diverse application scenarios and maximize privacy and usability (see I-C). This may encompass the incorporation of temporal information available in dynamic datasets (see I-D) through *Privacy-Preserving Temporal Record Linkage (PPTRL)* [20] techniques or the adoption of scalable linkage approaches [21], [22] in the context of distributed and heterogeneous sources (see I-A). Concerning the scalability issue, our proposal is the adaptation of advanced Blocking techniques designed by the DBGroup [23]–[25] leveraging the proposed framework to parallelize the generation of private blocks at the source side and the TTP to generate accurate candidate pseudonym pairs.

C. Data Fusion

Data Fusion is the process of fusing duplicate entries from different sources into a single unified record. It is aimed at increasing the conciseness and consistency of data that are made available to users and applications to facilitate data analysis. To this end, the outcome of PPDI is in plain format and therefore includes only the SPI as the QID possess the potential to enable re-identification. For this reason, the TTP may employ traditional data fusion approaches to merge plain SPI related to pseudonyms categorized as referencing the same real-world entity. Different high-level strategies exist to remove redundant data and select the best Resolution Function or policy to merge common attributes [26]. Numerous articles tend to make the assumption that databases generated through PPRL inherently possess a high level of security. Nevertheless, datasets that contain a rich repository of information pertaining to the same entity (see I-A) become more susceptible to re-identification attacks.

D. Privacy-Preserving Data Publishing and Warehousing

The subsequent phase to the PPDI process is *Privacy-Preserving Data Publishing* (PPDP) [27], with the goal of offering the Consumer Domain a privacy-preserving view of the unified data, restricting privacy disclosure and the potential of re-identification attacks. This problem is the central topic of statistical disclosure control [28], which aims to reduce the risk of information disclosure by restricting or generalizing the amount of data released. However, this reduces the usability of the data (see I-C) and eliminates hidden patterns. To address this limitation, we propose to additionally investigate data augmentation and imputation techniques [29] to achieve generalization and increase the fairness of the data for underrepresented subgroups (see I-A).

The crucial challenge, however, resides in the assessment and preservation of privacy, leading to the establishment of numerous research programs to pursue this objective proposing advanced privacy measurements [30]. Nevertheless, they tend to concentrate on PPRL, while the unaddressed aspects need the consideration of many different strategies, including PPDP [31]. From our perspective, PPDI and PPDP, which have often

been treated separately, need to be collectively approached, especially in addressing privacy. Another significant aspect pertains to the enduring maintenance of an integrated and secure database. Consequently, a pivotal decision consists of whether to virtualize the PPDI/PPDP process or to materialize its outcome in a Data Warehouse. This evaluation necessitates considering the dynamic evolution of datasets over time (see I-D), as this presents a substantial privacy risk by increasing adversaries' capacity to infer sensitive information across various releases. This entails the need to discern critical privacy aspects within data materialization and virtualization scenarios, and propose architectural privacy-preserving alternatives [32], [33]. It is worth noting that no research has systematically and globally addressed the assessment of privacy in relation to dynamic temporal aspects.

III. Conclusion and Future Works

This paper provides an overview of the Privacy-Preserving Data Integration (PPDI) process encompassing numerous challenges, especially in the context of Big Data, and related investigated solutions in the literature while recognizing the potential for many unexplored avenues. From our perspective, the broad spectrum of tasks and issues pertaining not only to Privacy-Preserving Data Integration but also to Privacy-Preserving Data Publishing advocates for a holistic approach to asset and preserve privacy. To this goal, we looked at issues that, to the best of our knowledge, have not received extensive coverage in the existing literature. Furthermore, we present an overview of the current state of our research efforts for the design of a novel comprehensive PPDI framework facing some of these issues. The lack of empirical metrics to quantify the trade-off between privacy and utility presents a significant challenge. To this end, a future direction is the study of specific privacy-loss metrics and approaches to facilitate the selection of tailored methods and techniques to address the challenges of diverse real-world application scenarios.

Acknowledgment

We wish to thank all the members of DBGroup.

References

[1] Chris Clifton. et al., "Privacy-preserving data integration and sharing," in *DMKD*. ACM, 2004, pp. 19–26.

[2] G. Pravettoni and S. Triberti, *P5 eHealth: An agenda for the health technologies of the future*. Springer Nature, 2020.

[3] Alfredo Cuzzocrea. et al., "Analytics over large-scale multidimensional data: The big data revolution!" in *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP*, 2011, p. 101–104.

[4] Sonia Bergamaschi. et al., "From data integration to big data integration," in *A Comprehensive Guide Through the Italian Database Research*, ser. Studies in Big Data. Springer International Publishing, 2018, vol. 31, pp. 43–59.

[5] Luca Bolognini. et al., "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.*, vol. 33, no. 2, pp. 171–181, 2017.

[6] Daochen Zha. et al., "Data-centric AI: perspectives and challenges," *CoRR*, vol. abs/2301.04819, 2023.

[7] Leon Willenborg. et al., *Elements of statistical disclosure control*. Springer Science & Business Media, 2012, vol. 155.

[8] Anushka Vidanage. et al., "Taxonomy of attacks on privacy-preserving record linkage," *J. Priv. Confidentiality*, vol. 12, no. 1, 2022.

[9] Lisa Trigiante. et al., "Privacy-preserving data integration for digital justice," in *International Conference on Conceptual Modeling*. Springer, 2023, pp. 172–177. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-47112-4_16

[10] ——, "Privacy-preserving data integration for health," *31st Symposium on Advanced Database Systems*, 2023. [Online]. Available: https://ceur-ws.org/Vol-3478/paper39.pdff

[11] Lisa Trigiante, "Analysis and experimentation of state-of-the-art privacy-preserving record linkage techniques in data integration environments," Master's thesis, Unimore, 2022. [Online]. Available: https://dbgroup.ing.unimore.it/publication/TrigianteL_Master_Thesis.pdf

[12] Alfredo Cuzzocrea. et al., "Privacy-preserving big data exchange: Models, issues, future research directions," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5081–5084.

[13] R. Schnell, "Privacy-preserving record linkage," in *Methodological Developments in Data Linkage*. John Wiley & Sons, 2015, pp. 201–225.

[14] Yehuda Lindell. et al., "Secure multiparty computation for privacy-preserving data mining," *J. Priv. Confidentiality*, vol. 1, no. 1, 2009.

[15] Sonia Bergamaschi. et al., "Data integration," in *Handbook of Conceptual Modeling*. Springer, 2011, pp. 441–476.

[16] Martin Franke. et al., "Post-processing methods for high quality privacy-preserving record linkage," in *DPM/CBT@ESORICS*, ser. Lecture Notes in Computer Science, vol. 11025. Springer, 2018, pp. 263–278.

[17] Murat Kantarcioglu. et al., "A cryptographic approach to securely share and query genomic sequences," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 5, pp. 606–617, 2008.

[18] Sirintra Vaiwsri. et al., "Accurate privacy-preserving record linkage for databases with missing values," *Information Systems*, vol. 106, p. 101959, 2022.

[19] Dinusha Vatsalan. et al., "A taxonomy of privacy-preserving record linkage techniques," *Inf. Syst.*, vol. 38, no. 6, pp. 946–969, 2013.

[20] Thilina Ranbaduge. et al., "Privacy-preserving temporal record linkage," in *IEEE International Conference on Data Mining, ICDM*. IEEE Computer Society, 2018, pp. 377–386.

[21] Dinusha Vatsalan. et al., "Incremental clustering techniques for multi-party privacy-preserving record linkage," *Data Knowl. Eng.*, vol. 128, p. 101809, 2020.

[22] A. Cuzzocrea, "Privacy and security of big data: Current challenges and future research perspectives," in *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data*, 2014, p. 45–47.

[23] Domenico Beneventano. et al., "*BLAST2*: An efficient technique for loose schema information extraction from heterogeneous big data sources," *ACM J. Data Inf. Qual.*, vol. 12, no. 4, pp. 18:1–18:22, 2020.

[24] George Papadakis. et al., "Three-dimensional entity resolution with jedai," *Inf. Syst.*, vol. 93, p. 101565, 2020.

[25] Giovanni Simonini. et al., "Progressive entity resolution with node embeddings," in *SEBD 2022*, ser. CEUR Workshop Proceedings, vol. 3194. CEUR-WS.org, 2022, pp. 52–60.

[26] Domenico Beneventano. et al., "Entity resolution and data fusion: An integrated approach," in *SEBD 2019*, ser. CEUR Workshop Proceedings, M. Mecella, G. Amato, and C. Gennaro, Eds., vol. 2400. CEUR-WS.org, 2019.

[27] Bee-Chung Chen. et al., "Privacy-preserving data publishing," *Foundations and Trends® in Databases*, vol. 2, no. 1–2, pp. 1–167, 2009.

[28] Michael Comerford. et al., "Statistical disclosure control : an interdisciplinary approach to the problem of balancing privacy risk and data utility," Ph.D. dissertation, University of Glasgow, UK, 2014.

[29] Fernando Moncada Martins. et al., "Data augmentation effects on highly imbalanced EEG datasets for automatic detection of photoparoxysmal responses," *Sensors*, vol. 23, no. 4, p. 2312, 2023.

[30] Anushka Vidanage. et al., "A vulnerability assessment framework for privacy-preserving record linkage," *ACM Transactions on Privacy and Security*, 2023.

[31] Benjamin C. M. Fung. et al., "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, 2010.

[32] Benjamin Fabian. et al., "Privacy-preserving data warehousing," *Int. J. Bus. Intell. Data Min.*, vol. 10, no. 4, pp. 297–336, 2015.

[33] Alfredo Cuzzocrea. et al., "Data warehousing and olap over big data: Current challenges and future research directions," in *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP*, 2013, p. 67–70.