

Privacy-Preserving Data Integration for Health: Adhering to OMOP-CDM Standard

Lisa Trigiante^{1,*}, Domenico Beneventano¹

¹University of Modena and Reggio Emilia, Modena, Italy

Abstract

The digital transformation of health processes and the resulting availability of vast amounts of health-related data about patients offer significant promise to advance multiple medical research projects and enhance both the public and private healthcare systems. Exploiting the full potential of this vision requires a unified representation of different autonomous data sources to facilitate detailed data analysis capacity. To this end, OMOP CDM has emerged as the de facto standard for organizing healthcare data from diverse sources. However, collecting and processing sensitive data about individuals leads to consideration of privacy requirements and confidentiality concerns. Privacy-Preserving Data Integration (PPDI) is the process of establishing a unified view of personal data across multiple data sources while protecting the privacy of individuals represented in the underlying data. This discussion paper offers a concise overview of the research field related to PPDI, highlighting associated challenges and opportunities within the healthcare domain. In particular, it delves into the specific research challenges encountered by the PPDI process alongside the utilization of OMOP-CDM, with particular attention directed towards the Schema Alignment phase and the classification of data based on identifiability and privacy.

1. Introduction

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, enabling users to access a unified and consistent view of diverse data sources. However, the integration of personal data is limited by ethical and privacy concerns. The European *General Data Protection Regulation* (GDPR) bases the classification of data content on the concepts of identifiability and privacy:



- *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).

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

✉ lisa.trigiante@unimore.it (L. Trigiante); domenico.beneventano@unimore.it (D. Beneventano)

ORCID 0000-0002-2021-9259 (L. Trigiante); 0000-0001-6616-1753 (D. Beneventano)

  © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

Privacy-Preserving Data Integration (PPDI) [1] is the process 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.

The GDPR leads toward the adoption of specific techniques [2] to prevent internal parties involved in the PPDI process and external adversaries from the possibility of identifying a specific individual, called *Re-identification*.

Our research in the field of PPDI has encompassed concrete application projects, such as the design and development of a Proof of Concept (PoC) for the *Criminal Data Warehouse* project, establishing a PPDI process across Italian legal data sources to assess the recidivism phenomena [3]. However, we intended from the design stage to accommodate different application scenarios and not tailor solutions to the justice domain. To this end, our collaboration with the Health Departments of the Emilia Romagna region have underscored the challenges inherent in privacy-preserving processing of health-related data [4, 5], emphasizing the necessity for a standard and comprehensive PPDI framework. Pursuing this line, the *European Health Data Evidence Network (EHDEN)* aims to promote the adoption of the *Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)* in Europe. In furtherance of this objective, our research group is participating in the European ARISTOTELES project, which includes a research strand focused on the PPDI process adherent to OMOP-CDM standards.

This discussion paper aims to delineate our current and future research efforts directed toward the creation of a novel and comprehensive PPDI framework within the EHDEN ecosystem.

- Section 2 provides a concise overview of the PPDI process devised for our framework (a more in-depth discussion was addressed in [6, 5]).
- In Section 3 we present the OMOP-CDM standard and discuss the major advantages and drawbacks, particularly concerning privacy issues that have not been adequately addressed in the literature.
- Section 4 proposes a primary contribution to overcome some of these challenges, exploiting a semantic-based tool to classify schema elements in QID and SPI, facilitating the schema alignment between local sources and OMOP-CDM and allowing to maximize the trade-off between privacy and utility.
- Finally, in Section 5, we conclude and provide insights for future directions and developments.

2. COMPREHENSIVE PPDI FRAMEWORK

In this section, we outline the methodology and architectural approach devised to support the creation of a novel and comprehensive PPDI framework. The idea behind the PPDI framework is an incremental extension of the *MOMIS (Mediator environment for Multiple Information Sources)* [7] Data Integration system toward a *Trusted Third-Party (TTP)* [8] 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.

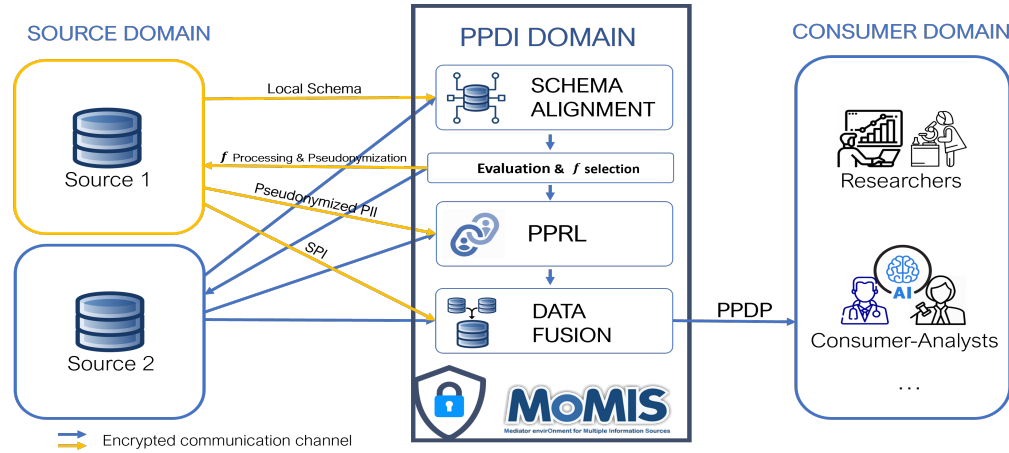


Figure 1: Schema of the PPDI Architecture.

2.1. Privacy-Preserving Data Integration process

The PPDI process usually 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. Furthermore, within the privacy context, this step involves the classification of data based on identifiability and privacy. The sets of PII and SPI are typically considered disjointed in a PPDI process and undergo distinct procedures.
- **Privacy-Preserving Record Linkage (PPRL)** resolves inconsistencies at the tuple level by identifying records about the same individuals from different sources. 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. To this end, PII undergoes specific pseudonymization techniques to facilitate record linkage while preventing re-identification.
Pseudonymization [9] is the process of replacing PII with a *pseudonym* (or encrypted code), to allow further processing.
- **Data Fusion** resolves inconsistencies at the value level by fusing duplicate entries from different sources and creating a unique record for each individual. Data Fusion 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 PII and QID possess the potential to enable re-identification.

2.2. Privacy and usability trade-off

In real-world privacy scenarios, with any information disclosure, there is always some privacy loss, and with any masking (or pseudonymization) technique, there is always some information loss. An important issue of privacy-preserving approaches 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) and to maximize privacy protection. For instance, one of the key dimensions for assessing the usefulness of data sharing is de-duplication (aka record linkage). On the other hand, the evaluation of privacy is one of the biggest impediments in a PPDI process 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 [10]. Some methods have been proposed to pursue this objective [11], nevertheless, they tend to concentrate only on PPRL. 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. From our perspective, PPDI needs to be collectively approached, especially in addressing privacy.

This discussion paper delves into the specific research challenges encountered by the PPDI process alongside the utilization of OMOP-CDM, with particular attention directed toward the privacy and usability trade-off.

3. OMOP Common Data Model

The advent of Big Health Data (BHD) has led to an upsurge in the need for methods to effectively manage their information content and offer a unified view to enable efficient analysis. The intrinsic aspects of BHD require careful consideration and impose strict demands on the data resulting from the PPDI process concerning completeness, consistency, interoperability, and scalability over time.

The *Observational Health Data Sciences and Informatics*¹ (OHDSI) program proposed the OMOP-CDM to standardize the structure and content of health data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP-CDM is the *OMOP standardized vocabularies* which allow organization and standardization of medical terms. OMOP-CDM plays a crucial role in addressing the challenges of data heterogeneity and interoperability among disparate healthcare systems by facilitating consistency, compatibility, and efficiency of the integration process. Moreover, OMOP-CDM addresses scalability challenges by accommodating large datasets and allowing for the independent addition of new sources, thereby empowering the management of vast amounts of health data with high performance and reliability. For these reasons, EHDEN launched a program aimed to promote the adoption of OMOP/OHDSI in Europe, addressing the challenges in generating insights and evidence from real-world clinical data on a large scale. The project's goal is to assist patients, clinicians, regulators, governments, and the industry in understanding well-being, disease, treatments, and outcomes, as well as new therapeutics and novel devices. Due to this initiative, the OMOP-CDM has been widely adopted across various healthcare systems, research institutions, and data

¹www.ohdsi.org

repositories worldwide, and now constitutes a vast repository of health data for observational studies and evidence-based research. The literature concerning the procedure to harmonize data with respect to OMOP-CDM encompasses a diverse range of data types, including but not limited to electronic health records (EHRs) [12], claims datasets [13], registries [14], and clinical trial [15]. Within such literature, the mapping process to ensure standardized representation and compatibility with OMOP CDM can be summarized in the following three phases:

1. *Vocabulary mapping*: is the process of mapping elements from a local data source (especially medical terms) to an appropriate standard concept defined within the OMOP vocabularies.
2. *Data tables mapping*: is the process of aligning the structure and semantics of the local data source with the standardized tables and fields defined in the OMOP-CDM.
3. *Extract-Transform-Load (ETL)*: is the process that involves the extraction of local data and their transformation based on the Data tables mapping rules defined in the previous steps. Finally, local data are loaded into the OMOP CDM-compliant database.

One of the main drawbacks, nevertheless, is that while in the numerous works that undertake health database mapping in OMOP [12, 13, 14], the data table mapping phase is often performed manually and/or coded only in the ETL stage. This negatively affects the tradeoff between privacy and usability of the overall process.

3.1. Privacy and usability trade-off in OMOP/OHDSI ecosystem

The vocabulary mapping process is extremely difficult, time-consuming, and mostly conducted manually by domain experts. To facilitate this human-in-the-loop process, some tools are provided by OHDSI. The most important one is Usagi², a vocabulary mapping tool that utilizes probabilistic algorithms to suggest mappings between local source terminologies and standard vocabularies to domain experts. One of the major drawbacks of Usagi is its exclusive reliance on a probabilistic algorithm based on syntactic matching. This results in limited accuracy, particularly with ambiguous terms and complex relationships, along with linguistic dependence, challenges in adapting to domain-specific vocabularies, and scalability issues. This aspect highly affects the utility of the resulting data. Different research programs have been established to improve Usagi's performance. Deep learning-based methods demonstrate to outperform both Usagi and previous simple word-level matching algorithms. However, the main limitation lies in the need for a conspicuous and accurate training set as the presence of negative training samples significantly affects the outcomes. Other researchers extended the vocabulary mapping to different languages [16] through automatic translation methods. However, challenges remain in translating nonstandard expressions and resolving abbreviations into full names. To overcome the limitations of these methods, we are developing techniques that will be briefly presented in Section 4.2. For instance, [17] discusses how to translate multilingual nonstandard expressions and resolve abbreviations.

The focus of this article is indeed on the privacy issues that arise when transforming data into OMOP-CDM. One of the main privacy challenges encountered by the PPDI process, alongside

²www.ohdsi.org/software-tools

the utilization of OMOP-CDM, concerns the fact that mapping large amounts of data to the OMOP-CDM raises significant concerns about protecting QID; as clinical terminologies expand to include new terms that may capture QID, institutions may inadvertently start using them in clinical data ETL processes. This can potentially put institutions and patients at risk if not addressed. The OHDSI consortium strongly cautions against this during the ETL process, as certain vocabularies may contain terms that represent phone numbers, emails, and other QID information, rather than clinical observations [18]. This highlights the importance of carefully considering the potential risks and implementing appropriate safeguards when mapping health data to the OMOP-CDM. To address this challenge, we envisaged a method to semi-automatically classify OMOP-CDM attributes into QID and SPI, presented in Section 4.1.

4. OMOP-CDM Privacy-Preserving Framework

The objective of this section is to present the methodology devised to develop additional privacy-preserving services to be implemented within our PPDI framework, based on the re-implementation of MOMIS, (see Section 2) to enhance the process of harmonizing data with OMOP-CDM standards (see Section 3). An open-source version of MOMIS is currently maintained by DataRiver³. DataRiver participated in various European projects aligning different health-related cross europe sources to OMOP-CDM and was certified as an SME by the IMI EHDEN consortium for its support to healthcare facilities in standardizing health data according to the OMOP CDM standard and providing additional services in the EHDEN and OHDSI ecosystem. This experience has proven important in identifying the main issues of OMOP/OHDSI approach in concrete applications. In light of these reasons, presented in Section 4.1, we are investigating techniques to semi-automatically classify QID and SPI and facilitate the schema alignment according to OMOP-CDM.

4.1. Classification of OMOP-CDM data

Classifying data according to identifiability in a real-world scenario is a complex task that highly impacts the overall trade-off between privacy and data usability. SPI and QID can overlap and the combination of attributes identifying an individual may vary from person to person depending on the rarity of attribute values.

To address this challenge, we intend to develop a tool adherent to OHDSI principles to semi-automatically classify OMOP-CDM attributes into QID and SPI. This tool applied to the OMOP-CDM attributes and to the terminologies in the main vocabularies, will allow us to determine whether they are QID or SPI.

This outcome can be leveraged not only in the mapping of new local sources to OMOP-CDM but also across all the previously mentioned systems and projects where the transformation into OMOP-CDM is already underway (see section 2), enabling validation of the ETL process from a privacy perspective. For instance, the local attributes mapped to OMOP concepts classified as identifiers will necessitate the consideration of appropriate controls and privacy protection techniques.

³originally founded as a spin-off of DBGroup dbgroup.unimore.it/

To investigate the feasibility of our idea and highlight potential issues within the process, we took as a starting point the already available manually annotated terminologies of PII related to medical data. As it is likely impossible to capture all codes that can potentially contain PII, due to their wide variability, an initial resource is presented in [18]⁴. Another one is the PPI (Participant Provided Information) terminology, a standard vocabulary in OMOP related to the 'All of Us' program, which focuses on collecting health data from a diverse group of participants [19]. However, the PPI terminology is not linked by any relationship to any other vocabulary. Furthermore, these resources are not directly linked to OMOP-CDM vocabularies, therefore we explored various procedures to match these resources with OMOP-CDM terminologies and leverage this knowledge to classify the mapped attributes as PII.

- Initially, we investigated different mapping techniques based on syntactic probabilistic algorithms, which as expected yielded poor results, but with the inclusion of specific data pre-processing approaches, they do tend to identify some potentially identifiable attributes.
- In addition, experiments were conducted exploiting Large Language Models (LLM). We tested some of the most renowned open-source LLM instructing them with specific examples to perform the mapping. Initially, some models have shown acceptable performance but later exhibited hallucinations regarding column names.
- Subsequently, we focused on techniques specifically developed for the classification and annotation tasks. For instance, we adapted the methodology used in [20] by defining identifiable and sensitive data classes as labels for annotation. However, the performance of this approach was limited by the impossibility of accessing plain-text column data.

In consideration of these experiments, we consider that a potential approach to achieving accurate classification methods may lie in training an instance of LLM specifically for this task. However, these models do not provide high reliability as they are also based on probabilistic syntactic principles and fail to capture the semantics of concepts [21]. Therefore, we believe that employing symbolic, explainable, and semantic-based methods will yield more promising results.

4.2. Data mapping to OMOP CDM

In the application case where a local source has not already been harmonized to OMOP-CDM, it is not only necessary to perform the data classification procedure described above, but also the Vocabulary Mapping, Data tables mapping and ELT process described in Section 2. This approach can be locally performed by a single source, participating in the PPI framework outlined in 1. However, the harmonization process is extremely challenging, needing to be accomplished quasi-totally manually by domain experts.

From a practical standpoint, the first steps of the harmonization process can be overlaid on the Schema Matching phase of data integration processes (Section 2). In the majority of data integration projects the schema matching phase is implemented following a bottom-up approach, finding the correspondences between the different schema of Local sources and producing a

⁴github.com/data2health/next-gen-data-sharing/blob/master/CodesWithPPIPotential.csv

unique integrated Global Schema. Within the OMOP/OHDSI ecosystem, the global schema is represented by OMOP-CDM. Therefore this phase is carried out using a top-down approach, aligning each local schema to OMOP-CDM and producing mapping rules to harmonize the original data. This allows parallelization across multiple local sources and the addition of new ones, dealing with scalability and interoperability issues (see Section 2) of the traditional bottom-up approach. However, within a privacy-preserving context, to prevent data privacy disclosure is not possible to access the original data in plain format, but only metadata, attribute names, and their associated descriptions, therefore only schema-level matching methods can be applied. The schema-level matching method SMAT[22] is considered a baseline within the OMOP-CDM and it is based on a deep learning model incorporating NLP techniques. Instead, within the PoC for the Criminal Data Warehouse project [3] we employed a schema-level matching method based on annotation.

4.2.1. Semantic Privacy-Preserving Schema Annotator

To date, the re-implementation of MOMIS has consisted of the development of different microservices to perform semantic-based schema annotation of property names and values, exploiting the semantic knowledge of multiple thesauri in multiple languages. The system allows scalability on multiple simultaneous pipelines and thesaurus configuration. Among them the OMOP standard for the medical field, including all the different standardized vocabularies (such as SNOMED, ICD10, and several others) and the OMW - Open Multilingual WordNet, offering good linguistic coverage. Furthermore 35 translation services are available, supporting all major languages and specific functions for the identification of non-canonic terms [17] are available allowing better performance. E.g., the term “consenso” is not found in OMOP, however the automatic translation “consent” can be correctly annotated. We will provide a more accurate description of this system in other publications as it is still in the development phase, but from the initial results, our method appears to achieve better results compared to SMAT [22].

On the other hand, it is also advisable to contemplate scenarios where accessing the local schema is unfeasible and hence explore the concept of *Privacy-Preserving Schema Matching* (PPSM) [23]. It establishes that at the time when the schema matching is conducted, no concrete information regarding the local data and schema has been released to the PPD framework. For this reason, future work will concern the application of PPSM methods [23] to the OMOP-CDM context, based on the classification and privacy assessment presented in 4.1.

5. Conclusion and Future Works

This paper provides an overview of the Privacy-Preserving Data Integration process encompassing numerous challenges, especially in the context of Big Health Data. Section 3 presents the OMOP-CDM standardized data structure, the process of harmonizing healthcare datasets to OMOP-CDM format and discussing the assessment of the privacy and usability trade-off of this process. In light of this, Section 4 presents a solution to address the main privacy issues in the context of OMOP-CDM. Namely, the design and development of specific methodologies and tools, adherent to the principles of OHDSI, to classify data according to identifiability and privacy, and perform a Privacy-Preserving Schema Matching [23] process according to

OMOP-CDM. This solution is still in the development phase but promises good results in terms of performance and privacy protection compared to other state-of-the-art systems. From our perspective, the broad spectrum of tasks and issues about the process of harmonizing data according to OMOP-CDM that, to the best of our knowledge, have not received extensive coverage in the existing literature allow for many future developments. One of the main shortcomings of this process that limits its usefulness is the absence of de-duplication [24]. To this end, a possible future development is the adaptation of Privacy-Preserving Record Linkage techniques and Data fusion approaches to perform de-duplication within the OMOP-CDM context.

Acknowledgment

We wish to thank all the members of DBGroup. Lisa Trigiante wishes to mention that her PhD project is funded by MIUR under D.M.351 with the Emilia Romagna region as partner.

References

- [1] Chris Clifton. et al., Privacy-preserving data integration and sharing, in: DMKD, ACM, 2004, pp. 19–26.
- [2] 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.* 33 (2017) 171–181.
- [3] Lisa Trigiante. et al., Privacy-preserving data integration for digital justice, in: *International Conference on Conceptual Modeling*, Springer, 2023, pp. 172–177. URL: https://link.springer.com/chapter/10.1007/978-3-031-47112-4_16.
- [4] Lisa Trigiante. et al., Privacy-preserving data integration for health, *31st Symposium on Advanced Database Systems (2023)*. URL: <https://ceur-ws.org/Vol-3478/paper39.pdf>.
- [5] L. Trigiante, D. Beneventano, S. Bergamaschi, [vision paper] privacy-preserving data integration, in: *2023 IEEE International Conference on Big Data (BigData)*, IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 5614–5618. URL: <https://doi.ieeecomputersociety.org/10.1109/BigData59044.2023.10386703>. doi:10.1109/BigData59044.2023.10386703.
- [6] Lisa Trigiante, Analysis and experimentation of State-of-the-Art Privacy-Preserving Record Linkage techniques in Data Integration environments, Master’s thesis, Unimore, 2022. URL: https://dbgroup.ing.unimore.it/publication/TrigianteL_Master_Thesis.pdf.
- [7] Sonia Bergamaschi. et al., Data integration, in: *Handbook of Conceptual Modeling*, Springer, 2011, pp. 441–476.
- [8] R. Schnell, Privacy-preserving record linkage, in: *Methodological Developments in Data Linkage*, John Wiley & Sons, 2015, pp. 201–225.
- [9] Daochen Zha. et al., Data-centric AI: perspectives and challenges, *CoRR* abs/2301.04819 (2023).
- [10] Anushka Vidanage. et al., Taxonomy of attacks on privacy-preserving record linkage, *J. Priv. Confidentiality* 12 (2022).
- [11] Anushka Vidanage. et al., A vulnerability assessment framework for privacy-preserving record linkage, *ACM Transactions on Privacy and Security* (2023).

- [12] A. Matcho, P. Ryan, D. Fife, C. Reich, Fidelity assessment of a clinical practice research datalink conversion to the omop common data model, *Drug safety* 37 (2014) 945–959.
- [13] A. Haberson, C. Rinner, A. Schöberl, W. Gall, Feasibility of mapping austrian health claims data to the OMOP common data model, *J. Medical Syst.* 43 (2019) 314:1–314:5. URL: <https://doi.org/10.1007/s10916-019-1436-9>. doi:10.1007/S10916-019-1436-9.
- [14] M. Y. Garza, G. D. Fiol, J. D. Tenenbaum, A. Walden, M. Nahm, Evaluating common data models for use with a longitudinal community registry, *J. Biomed. Informatics* 64 (2016) 333–341. URL: <https://doi.org/10.1016/j.jbi.2016.10.016>. doi:10.1016/J.JBI.2016.10.016.
- [15] H. Liu, S. Carini, Z. Chen, S. P. Hey, I. Sim, C. Weng, Ontology-based categorization of clinical studies by their conditions, *J. Biomed. Informatics* 135 (2022) 104235. URL: <https://doi.org/10.1016/j.jbi.2022.104235>. doi:10.1016/J.JBI.2022.104235.
- [16] A. Chechulina, J. Carus, P. Breitfeld, C. Gundler, H. Hees, R. Twerenbold, S. Blankenberg, F. Ückert, S. Nürnberg, Semi-automated mapping of german study data concepts to an english common data model, *Applied Sciences* 13 (2023). URL: <https://www.mdpi.com/2076-3417/13/14/8159>. doi:10.3390/app13148159.
- [17] D. Beneventano, S. Bergamaschi, S. Sorrentino, Extending wordnet with compound nouns for semi-automatic annotation in data integration systems, in: *Proceedings of the 5th International Conference on Natural Language Processing and Knowledge Engineering, NLPKE 2009, Dalian, China, September 24-27, 2009, IEEE, 2009*, pp. 1–8. URL: <https://doi.org/10.1109/NLPKE.2009.5313842>. doi:10.1109/NLPKE.2009.5313842.
- [18] E. R. Pfaff, M. A. Haendel, K. Kostka, A. Lee, E. Niehaus, M. B. Palchuk, K. Walters, C. G. Chute, Ensuring a safe (r) harbor: Excising personally identifiable information from structured electronic health record data, *Journal of Clinical and Translational Science* 6 (2022) e10.
- [19] A. of Us Research Program Investigators, The “all of us” research program, *New England Journal of Medicine* 381 (2019) 668–676.
- [20] K. Korini, C. Bizer, Column type annotation using chatgpt, *arXiv preprint arXiv:2306.00745* (2023).
- [21] W. Saba, Stochastic LLMs do not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs, 2023, pp. 3–19. doi:10.1007/978-3-031-47262-6_1.
- [22] J. Zhang, B. Shin, J. D. Choi, J. C. Ho, SMAT: an attention-based deep learning solution to the automation of schema matching, in: L. Bellatreche, M. Dumas, P. Karras, R. Matulevicius (Eds.), *Advances in Databases and Information Systems - 25th European Conference, ADBIS 2021, Tartu, Estonia, August 24-26, 2021, Proceedings*, volume 12843 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 260–274. URL: https://doi.org/10.1007/978-3-030-82472-3_19. doi:10.1007/978-3-030-82472-3_19.
- [23] T. Amagasa, F. Zhang, J. Sakuma, H. Kitagawa, A scheme for privacy-preserving ontology mapping, in: *Proceedings of the 18th International Database Engineering & Applications Symposium, IDEAS '14, Association for Computing Machinery, New York, NY, USA, 2014*, p. 87–95. URL: <https://doi.org/10.1145/2628194.2628232>. doi:10.1145/2628194.2628232.
- [24] F. N. Wirth, T. Meurers, M. Johns, F. Prasser, Privacy-preserving data sharing infrastructures for medical research: systematization and comparison, *BMC Medical Informatics Decis. Mak.* 21 (2021) 242. URL: <https://doi.org/10.1186/s12911-021-01602-x>. doi:10.1186/S12911-021-01602-X.