Anonymization Ontology: A Privacy-Preserving Framework for GDPR-Compliant Anonymization of Personal Health Data

Noopur Rai[1], Marta Dembska[1] and Sirko Schindler[1]

[1]German Aerospace Center (DLR), Institute of Data Science, Jena, Germany

Abstract

Medical research can immensely benefit from the use of personal health data. However, a large amount of health data remains unexplored due to existing data protection laws guaranteeing individual full control of their data. For a broad use in research, this health data needs to be anonymized without losing its specific characteristics. The anonymization ensures compliance with the data protection laws while providing a wealth of data, e.g., to train machine learning algorithms. In this paper, we present an ontology to describe generic anonymization software for GDPR-compliant health data anonymization as well as its characteristics and requirements.

Keywords

Data anonymization, Privacy methods, Anonymization software, GDPR, Ontology

1. Introduction

The General Data Protection Regulation (GDPR)[1] protects sensitive health data from potential misuse and risk of reidentification by requiring data to be fully anonymous so that the individual is no longer identifiable. Health data can potentially be linked back to the individual and, hence, might reveal sensitive information that should be kept confidential. Within the Avatar project[2], we aim to anonymize health data to make it GDPR-compliant, enabling its use for research without compromising individuals' privacy. Besides reuse of existing work like the ARX Data Anonymization Tool [1] and development of new anonymization techniques, this also includes search and retrieval of health data that is properly anonymized. Here, our approach is based on semantic descriptions of data, anonymization techniques, and user needs. In this paper, we introduce an anonymization ontology that semantically describes the anonymization software. We combine concepts of privacy methods with GDPR requirements, and we model characteristics and capabilities of anonymization software.

2. Anonymization Ontology

The anonymization ontology conceptualizes generic anonymization software to achieve GDPR-compliant anonymization. We semantically model the anonymization process as a modular software design to allow easy integration of new privacy methods and data from diverse data domains. Figure 1 illustrates the relationships between key classes and properties of the anonymization ontology. It can be structured as follows: The entry point is represented by the software itself as well as its components whose details can be described using different `Modules` *(red box; (1)Software; upper left)*. A first module considers the dataset a software is applicable to *(green box; (2)Dataset; left)*. Here, restrictions like the datasets' inner structure and processable datatypes can be specified. Another module *(blue box; (3)Anonymization; right)* allows to describe the underlying anonymization process, transformation and

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025

Noopur.Rai@dlr.de (N. Rai); Marta.Dembska@dlr.de (M. Dembska); Sirko.Schindler@dlr.de (S. Schindler)

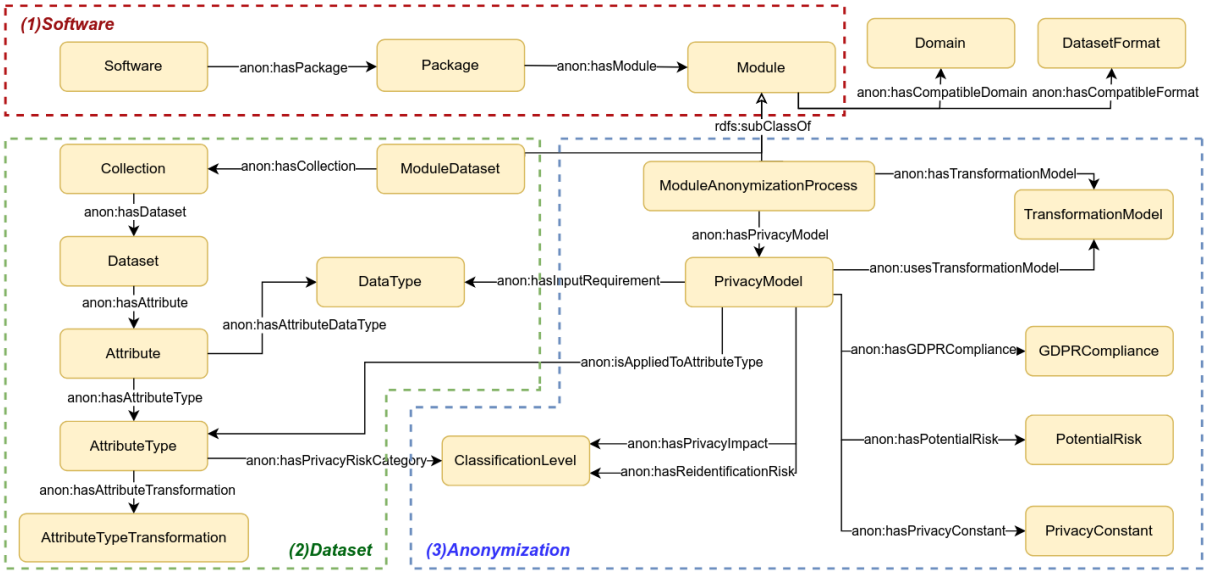
0009-0009-8214-190X (N. Rai); 0000-0002-8180-1525 (M. Dembska); 0000-0002-0964-4457 (S. Schindler)

[1]https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/

[2]https://www.avatar-projekt.de/en/

privacy models. Privacy models utilize transformation models such as generalization and suppression to transform a dataset to achieve anonymization.

Figure 1: An overview of the Anonymization Ontology. Classes omitted to readability's sake.



Can and Usenmez proposed an ontology-based, personalized anonymization model for privacy methods [3][2]. They model actual data, privacy methods, and the resulting anonymized data. In contrast, we focus on the anonymization process. We also allow to describe suitable data characteristics a specific approach can be used for. A specific approach may, e.g., be intended specifically for EEG (electroencephalogram) data or require a specific data structure as input. Finally, we conceptualize privacy methods, their risks, and GDPR compliance, thus enhancing the previous work. This detailed modeling will help in selecting suitable anonymization methods for specific datasets and research use cases in accordance with the Avatar project's goals.

The Anonymization ontology[4] is publicly available on Git repository[5] [3]. We are committed to further development and welcome comments, suggestions, and contributions via this repository.

Acknowledgments

This research is conducted as part of the Avatar project, funded by the German Federal Ministry of Education and Research (BMBF) under grant number *16KISA015*. We thank Thomas Köllmer, Cord Spreckelsen, Mark Hoffmann, and Patrick Aichroth for the discussions on privacy models and anonymization. We also acknowledge Jan Martin Keil for his valuable feedback on the anonymization ontology.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

[3] K-Anonymity, L-Diversity, Differential Privacy, and T-closeness
[4] https://w3id.org/AnonymizationOntology
[5] https://gitlab.com/dlr-dw/anonymization-ontology

References

[1] F. Prasser, F. Kohlmayer, Putting statistical disclosure control into practice: The ARX data anonymization tool, in: Medical Data Privacy Handbook, Springer International Publishing, Cham, 2015, pp. 111–148.

[2] O. Can, B. Usenmez, An ontology based personalized privacy preservation, in: Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, SCITEPRESS - Science and Technology Publications, 2019.

[3] N. Rai, Anonymization ontology: A privacy-preserving framework for GDPR-compliant anonymization of personal health data, 2024.