3rd NFDI4Energy Conference: Defining Energy Research Data Management Together
Proceedings to the 3rd NFDI4Energy Conference - Full Papers
https://doi.org/10.52825/ocp.v9i.3314

Published: 23 Mar. 2026

Collaborative Data Anonymization Framework With Energy Industry

Yuting Gao[1,*], Zhiyu Pan[1], and Antonello Monti[1,2]

[1]RWTH Aachen University, Germany
[2]Fraunhofer FIT, Germany
*Correspondence: Yuting Gao, yuting.gao@eonerc.rwth-aachen.de

Abstract. Data anonymization is essential for data privacy, enabling organizations, especially in the energy sector with IoT data, to comply with regulations while using sensitive information. Despite many existing anonymization methods and tools, it lacks clarity on integrating these tools into a complete process and ensuring effective collaboration with data providers. To overcome these gaps, this paper proposes a collaborative data anonymization framework that efficiently chains open-source tools to streamline the process and improve data provider involvement. The findings indicates while these tools can detect sensitive information and generate anonymized data, there are still limitations in metadata detection.

Keywords: Anonymization, Synthetic Data, Data Sharing, Energy Data Management

1. Introduction

In the energy domain, data is distributed across various stakeholders, including energy providers, grid operators, and consumers. A data sharing platform can facilitate data exchange by providing a central point for collecting, processing, and sharing energy-related data among different actors [1]. Due to privacy protection regulations and business interests, it is necessary to anonymize the data during data exchange. Moreover, ethical and security concerns pose additional challenges to data sharing [2]. Therefore, anonymization becomes an essential component of data contribution and access processes [3].

Data anonymization is applied to remove or transform personally identifiable information to prevent re-identification. In addition to traditional anonymization methods, e.g. , k-anonymity, l-diversity, or t-closeness, synthetic data has gained increasing attention [4] [5]. Synthetic data can serve as a form of anonymization if it prevents any link to real records. However, integrating existing tools into a complete process and collaborating effectively with data providers remains a challenge. To address these problems, this paper proposes a collaborative anonymization framework using open source tools.

The remainder of the paper is organized as follows: In Section 2, we discuss related work. Then, we present our framework in Section 3, and the evaluation in Section 4. Finally, we conclude our work in Section 5.

2. Related Work

2.1 Synthetic Data Generation

Synthetic data generation has been widely used to address data privacy. For synthetic data generation, there are two main imputation methods [6]: statistical and deep-learning methods. Statistical methods generate synthetic data by modeling the distribution of real data and then sampling from it. These methods typically involve two steps: parameter estimation and sampling. Techniques such as maximum likelihood estimation (MLE) or kernel density estimation (KDE) are used to determine statistical parameters like mean, variance, and covariance [7].

Deep learning models learn complex data distributions using neural networks. Typical approaches are Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs are probabilistic generative models that consist of an encoder and a decoder [8]. The encoder maps the input data into a latent space modeled as a multivariate distribution, typically a Gaussian, and the decoder samples from this latent space to reconstruct data samples. A GAN [9] consists of two networks: a generator, which creates synthetic samples from random noise, and a discriminator, which evaluates whether the samples are real or generated. GANs are trained through an adversarial process in which the generator creates synthetic data. The discriminator evaluates them to distinguish real samples from generated samples. Through this competition, the generator gradually learns to produce data that closely resembles real data distributions.

Building upon these statistical and deep learning foundations, more comprehensive frameworks have been developed to streamline the synthetic data generation pipeline. Synthetic Data Vault (SDV) [10] offers a more comprehensive and flexible solution to synthetic data generation. SDV includes deep generative models, probabilistic modeling, metadata management, constraint handling, and built-in evaluation metrics.

While these methods provide general-purpose solutions, their application to domain-specific problems often requires further adaptation. Within the energy domain, VAEs and GANs have been adapted to better handle time-series data. TimeGAN incorporates temporal patterns to generate realistic sequential energy data. Recurrent Conditional GANs and Conditional Wasserstein GANs have been employed to model energy consumption patterns. Moreover, hybrid models combining VAEs and GANs, known as VAE-GANs, have also been applied to energy data. These models use VAEs to learn the underlying data structure and GANs to improve sample realism, thereby producing synthetic time-series data that more closely reflect real energy patterns. The synthetic data are further leveraged to support load forecasting, renewable energy modeling, and power system simulation.

2.2 Data Anonymization Framework

In the energy domain, data are typically generated and managed by multiple actors, each of them may have distinct operational priorities and risk considerations. Therefore, collaboration with industry partners in anonymization decision-making is essential to ensure that both privacy requirements and domain-specific constraints are appropriately addressed.

A framework proposed in [11] leverages Relative Functional Dependencies (RFDs) to derive anonymization strategies that minimize utility loss, with Pareto-optimality used to identify optimal trade-offs between privacy and utility. The framework also considers the involvement of data owners in selecting generalization rules. However, it does not

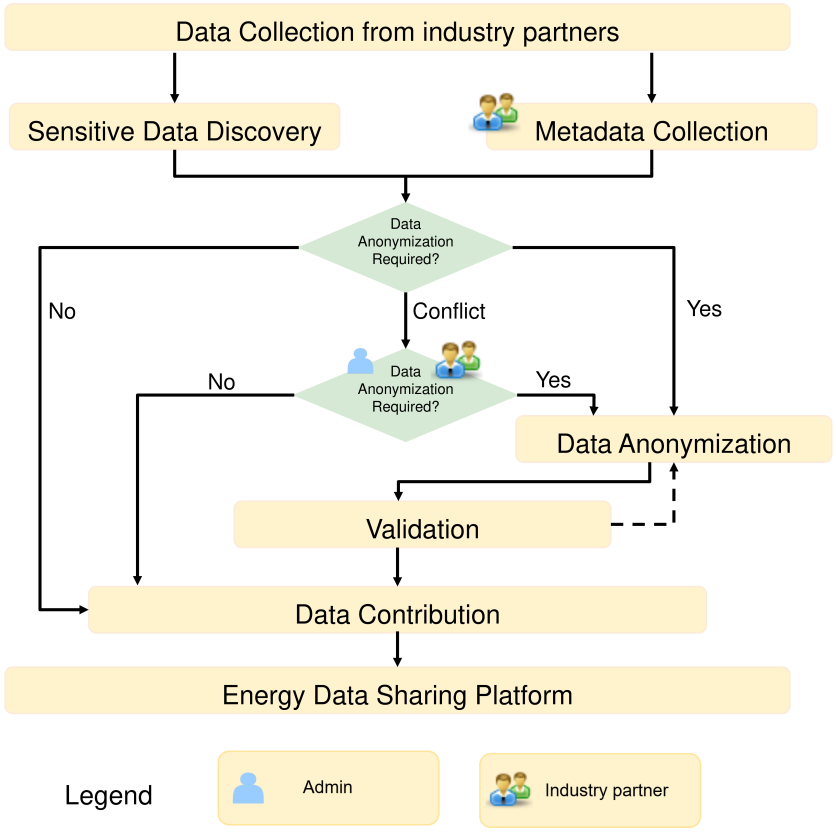

Figure 1. *Collaborative Data Anonymization Framework.*

incorporate metadata extraction and also lacks a validation mechanism for verifying data utility and privacy. These gaps highlight the need for integrated solutions, especially in synthetic data generation, which has increasingly evolved from isolated modeling techniques to integrated tool-chains that combine multiple components, including sensitive data discovery, privacy assessment, and quality validation [12][10] [13].

3. Methodology

This section is dedicated to describing the Collaborative Data Anonymization framework. In the following subsections, we first present an overview of the framework, outlining the tools used within the framework, followed by a detailed description of its implementation.

3.1 Framework Overview

In this subsection, we detail the proposed framework [14] for collaborative data anonymization. The framework, as shown in Fig.1, consists of several steps, including sensitive information detection, metadata collection, supportive decision-making on anonymization, data anonymization, and validation. These steps ensure that the shared dataset is privacy-safe and enable industry partners to contribute data securely.

Sensitive Data Discovery In this step, the aim is to automatically determine whether the tabular dataset provided by industry partners contains any sensitive information.

Metadata Collection In this step, metadata is collected from industry partners. In some cases, some datasets may have been anonymized by other collaborators. Integrating

this step into the framework can reduce redundant anonymization efforts. Moreover, it helps preserve domain-specific transformations already applied and accelerates the data contribution process.

Anonymization Decision In this step, the results of Sensitive Data Discovery are compared with the metadata provided by the industry partner to determine whether anonymization is required. When the two are consistent, there are two possibilities: if no sensitive information is detected, the data is uploaded directly to the Energy Data Sharing Platform; if sensitive content is identified, it proceeds to Data Anonymization. When a discrepancy arises between the automated detection results and the partner's metadata, the admin and the industry partner coordinate to resolve the conflict and agree on the appropriate course of action.

Data Anonymization In this step, the data is anonymized according to the previously made decisions.

Validation In this step, the anonymizated data is verify that privacy requirements are met and it complies with the requirements of the industry partner. If these criteria are not satisfied, the data anonymization process can be repeated.

Data Contribution In this step, the dataset is structured and anonymized. The step enables industry partners to preview the resulting synthetic data and select the preferred one for the final dataset.

Energy Data Sharing Platform The final step involves uploading the data to the energy data sharing platform in accordance with the FAIR principles, which ensure shared data can be easily located, accessed by authorized users, integrated with other datasets, and reused for various research purposes [1].

3.2 Implementation Tools

Presidio [15] is an open-source tool designed for detecting and anonymizing sensitive data, especially personally identifiable information (PII). It employs a hybrid approach to PII detection, combining rule-based pattern recognizers for structured data with NLP-based named entity recognition models for unstructured text. Each detected entity is assigned a confidence score, and the results from multiple recognizers are aggregated to ensure robust and accurate identification of sensitive information. The approach provides mechanisms for sensitive data identification and anonymization. However, it is limited in metadata management and lacks validation.

SDV provides several synthesizers for generating synthetic data using different approaches. The GaussianCopulaSynthesizer models the statistical distributions of each feature and their relationships, generating data that preserves basic statistical properties. The CTGANSynthesizer employs a GAN-based approach to capture more complex, nonlinear patterns in the data. Other synthesizers, such as TVAESynthesizer and CopulaGANSynthesizer, combine latent representations or copula models with adversarial training to handle more complex dependencies. These options enable SDV to generate synthetic data suitable for diverse tabular datasets. However, SDV has limitations in detecting metadata and evaluating sensitive information across domains. It does not provide a clear framework for collaborating on defining datasets that require anonymization.

3.3 Framework Implementation

Sensitive Data Discovery Presidio [15] is used here to detect PII type. The detected PII types and their confidence scores are used to support anonymization decisions.

Metadata Collection Relevant metadata are collected from the industry partner, including details on whether the dataset has been anonymized.

Anonymization Decision This is achieved by inspecting the metadata of the dataset based on SDV. Each column is associated with an *sdtype*. The *sdtype* in terms of personal info, e.g., email, name, and ssn, are flagged and presented to the industry partner, which enables them to verify the correctness of the detection and decide whether anonymization should be applied. This metadata-driven step enhances the collaboration with industry partners by providing sensitivity assessments and enabling informed decision-making regarding the need for anonymization.

Data Anonymization This step is implemented by using the SDV framework. SDV determines *sdtype* of each column based on automatic detection or user-provided metadata. User-provided metadata should include the column name and its data type. PII columns should be marked with pii $=$ True. When metadata is provided, it is used for later decisions that require an anonymization process. If metadata is not provided, the columns are automatically detected and classified into various *sdtype*. GaussianNormalizer is applied in this step to generate synthetic data [16].

Validation Synthetic data is evaluated by overall score [10], which is based on Column Shapes and Column Pair Trends.

4. Evaluation

In this section, we first introduce the datasets and evaluation metrics employed in our study, followed by a description of the experimental setup. The final subsection presents the experimental results and analysis.

4.1 Datasets

The datasets we use are collected from industry partners in the energy domain. These datasets are structured as a single tabular form and include various types of energy-related data, energy consumption data, and measurement types. The datasets cover three locations over a 12-month period.

4.2 Metrics

To evaluate the quality of synthetic data generation results, we use Overall Score of SDMetrics [10]. The Overall Score is the average of the Column Shapes and Column Pair Trends scores. Column Shapes is measured by KSComplement [17] for numerical and datetime columns, and TVComplement [18] for categorical or boolean columns, which quantify how closely each column's distribution in the synthetic data matches the original, with scores ranging from 0 (very different) to 1 (identical). Column Pair Trends is measured by CorrelationSimilarity [19] and ContingencySimilarity [20].

Table 1. *Overall Score of datasets.*

Data Type	Overall Score
Energy Consumption Data	0.773
Building Energy Management System Data	0.736
Renewable Energy Data	0.699
Sensor Data	0.809

Table 2. *Comparison of Original and Anonymized Energy Consumption Data.*

Month		Usage [kWh]		Cost [net in PLN]		Cost [gross in PLN]	
Orig.	Anon.	Orig.	Anon.	Orig.	Anon.	Orig.	Anon.
2017-01	01.05.2017	6561.00	3866.64	5148.17	4093.31	6332.24	5035.33
2017-02	01.05.2017	6947.88	4193.04	5092.37	4107.27	6263.63	5052.65
2017-03	01.04.2017	7254.04	6367.04	5282.47	5235.38	6497.43	6439.61
2017-04	01.08.2017	5594.92	4623.98	4917.50	4111.29	6048.52	5057.72
2017-05	01.08.2017	5059.80	4483.14	4808.05	4392.15	5913.91	5402.93
2017-06	01.07.2017	4539.12	6867.40	4339.91	5082.43	5338.09	6251.94
2017-07	01.08.2017	3436.48	3555.46	3892.64	3892.64	4787.95	4787.95
2017-08	01.08.2017	3876.28	6053.11	4071.57	4721.60	5008.03	5808.44
2017-09	01.08.2017	5931.76	3943.73	4905.81	4330.11	6034.15	5326.54

Orig. = Original data; Anon. = Anonymized data

4.3 Evaluation Setup

The input of our evaluation is data collection from industry partners. The original data are first transformed into a unified CSV format. Presidio is then applied to identify sensitive information, namely, Personally Identifiable Information (PII), in the dataset. The detection results are compared with the metadata provided by the industrial partner, which indicates whether the data has been anonymized.

All data are assumed to require anonymization to evaluate anonymization results across different data types. We also consider a typical conflict case in which Presidio does not recognize the location, e.g., Atlantis waterpark in energy consumption data as PII, while the industrial partner regards this as sensitive. In such cases, the metadata are updated based on the industry partner's decision. The column is marked as PII and included in the anonymization process.

4.4 Results

To evaluate the quality of synthetic data, we classify the tested datasets into four major types: Energy consumption data, Building energy management system data, Renewable energy data, and Sensor data. Table 1 presents the average overall scores for each data type.

In general, the results indicate that the GaussianCopulaSynthesizer performs well across various data types with overall scores exceeding 0.69. In particular, sensor and energy consumption data achieve relatively high scores of 0.809 and 0.773, respectively. In contrast, the renewable energy data shows a lower overall score of 0.699. This data type mainly contains photovoltaic data, including power generation from PV panels and solar irradiance. The underlying data patterns are more nonlinear, which makes it more difficult for general models to handle effectively.

Furthermore, we examine the Energy consumption data in detail and compare the original datasets with the anonymized results. The comparison shows that the time

Table 3. *Comparison of Original and Anonymized Location.*

Timestamp		Location	
Orig.	Anon.	Orig.	Anon.
21.09.2021 19:11	15.09.2021 17:11	Atlantis waterpark	Salinas-Jones
21.09.2021 19:41	14.08.2021 22:37	Atlantis waterpark	Burke Ltd
21.09.2021 19:41	19.08.2021 00:39	Atlantis waterpark	Duffy-Gonzalez
21.09.2021 19:41	02.09.2021 01:38	Atlantis waterpark	Rodriguez-Henderson
21.09.2021 19:41	14.08.2021 07:29	Atlantis waterpark	Moore, Saunders and Anderson
21.09.2021 20:11	09.08.2021 16:02	Atlantis waterpark	Warren Ltd
21.09.2021 20:11	20.08.2021 23:41	Atlantis waterpark	Brown Group
21.09.2021 20:11	03.09.2021 02:50	Atlantis waterpark	Erickson-Wilkerson
21.09.2021 20:11	16.08.2021 02:03	Atlantis waterpark	Wilson, Paul and Ramsey
21.09.2021 20:41	20.09.2021 19:56	Atlantis waterpark	Adams, Wong and Bradley

Orig. = Original data; Anon. = Anonymized data

structural patterns of the original data are not preserved after anonymization. In Table 2, the original Month column represents monthly time intervals, whereas the month is anonymized as calendar days. This change disrupts the temporal structure of the data. The anonymized table does not reflect the intended monthly aggregation. This occurs because the metadata process identified the column as a datetime without specifying its temporal hierarchy.

In addition to the temporal inconsistencies, we observe a limitation related to the identification of sensitive information in the energy datasets. During automatic metadata detection, location-related columns that do not contain street-level details are typically not classified as sensitive. Within our framework, we can manually set this as PII by adjusting the metadata definitions. As shown in Table 3, we marked Location as requiring anonymization, which resulted in the generation of a random address in the anonymized output. The original dataset comprises energy data for a single, specific location. Replacing the original single location with multiple addresses changes the spatial meaning of the data.

5. Conclusion

In this work, we followed the previously established framework for collaborative data anonymization process, implemented it by applying existing open-source tools Presidio and SDV for synthetic data generation, and evaluated the framework using datasets from the energy domain. The proposed framework supports column classification, PII detection, and collaboration with industry partners through metadata collection, enabling a systematic anonymization process. Based on the evaluation results, some information in the energy domain, such as location in the energy consumption dataset, is considered sensitive. However, these data are not always accurately detected during the automated anonymization. Furthermore, while synthetic data can effectively support privacy preservation, maintaining the statistical patterns of original energy datasets remains challenging.

For future work, we will focus on enhancing the framework with models designed for time-series energy data. This may include integrating or adapting models such as TimeGan. At the same time, improving the accuracy of sensitive information detection in energy datasets provides guarantee to privacy and data quality. In the current study, SDV models are used with their default settings, thus we plan to apply hyperparameter tuning and sensitivity analysis to assess their effect on the results. Another future direction is

to extend the evaluation in collaboration with industry partners, leveraging real-world datasets to validate the framework's effectiveness and applicability.

Data availability statement

Datasets are available from Zenodo https://doi.org/10.5281/zenodo.15270008.

Author contributions

Conceptualization, Y.G., Z.P., methodology, Y.G., Z.P.; writing—original draft preparation, Y.G; writing—review and editing, Y.G., Z.P., supervision, A.M..

Competing interests

The authors declare that they have no competing interests.

Funding

The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 501865131.

References

[1] Z. Pan, G. Gürses-Tran, C. Speck, P. Jaquart, M. Niebisch, and A. Monti, "Transparency and involvement of the energy-related industry in a data sharing platform", in *Proceedings of the Conference on Research Data Infrastructure*, vol. 1, 2023.

[2] J. K. Author, "Name of paper", *Title of Journal*, vol. X, no. Y, pp–pp, Dec. 2000. DOI: 10.....

[3] Z. Pan et al., *Process for contributing and accessing fair data*, Aug. 2025. DOI: 10.5281/zenodo.16735836. [Online]. Available: https://doi.org/10.5281/zenodo.16735836.

[4] A. Majeed, "Attribute-centric and synthetic data based privacy preserving methods: A systematic review", *Journal of Cybersecurity and Privacy*, vol. 3, no. 3, pp. 638–661, 2023.

[5] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data", *arXiv preprint arXiv:2211.10459*, 2022.

[6] M. Goyal and Q. H. Mahmoud, "A systematic review of synthetic data generation techniques using generative ai", *Electronics*, vol. 13, no. 17, p. 3509, 2024.

[7] G. Soltana, M. Sabetzadeh, and L. C. Briand, "Synthetic data generation for statistical testing", in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2017, pp. 872–882.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *arXiv preprint arXiv:1312.6114*, 2013.

[9] I. Goodfellow et al., "Generative adversarial networks", *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[10] *Synthetic data metrics*, Version 0.23.0, DataCebo, Inc., Aug. 2025. [Online]. Available: https://docs.sdv.dev/sdmetrics/.

[11] L. Caruccio, D. Desiato, G. Polese, G. Tortora, and N. Zannone, "A decision-support framework for data anonymization with application to machine learning processes", *Information Sciences*, vol. 613, pp. 1–32, 2022.

[12] B. Kaabachi et al., "A scoping review of privacy and utility metrics in medical synthetic data", *NPJ digital medicine*, vol. 8, no. 1, p. 60, 2025.

[13] Z. Qian, B.-C. Cebere, and M. van der Schaar, "Synthcity: Facilitating innovative use cases of synthetic data in different data modalities", *arXiv preprint arXiv:2301.07573*, 2023.
[14] Y. Gao, Z. Pan, S. Foroogh, and A. Monti, *Collaborative data anonymization process*, Aug. 2025. DOI: 10.5281/zenodo.16735882. [Online]. Available: https://doi.org/10.5281/zenodo.16735882.
[15] O. Mendels, C. Peled, N. Vaisman Levy, S. Hart, T. Rosenthal, L. Lahiani, et al., *Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images*, Microsoft, 2018. [Online]. Available: https://microsoft.github.io/presidio.
[16] SDV Developers. "Numerical transformers". Accessed: 2025-11-25. [Online]. Available: https://docs.sdv.dev/rdt/transformers-glossary/numerical.
[17] SDMetrics Documentation, *Kscomplement metric — sdmetrics*, Accessed: 2026-02-27, 2025. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/quality-metrics/kscomplement.
[18] SDMetrics Documentation, *Tvcomplement metric — sdmetrics*, Accessed: 2026-02-27, 2025. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/tvcomplement.
[19] SDMetrics Documentation, *Correlationsimilarity metric — sdmetrics*, Accessed: 2026-02-27, 2025. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/quality-metrics/correlationsimilarity.
[20] SDMetrics Documentation, *Contingencysimilarity metric — sdmetrics*, Accessed: 2026-02-27, 2025. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/quality-metrics/contingencysimilarity.