DATA ANONYMIZATION IN WASTE MANAGEMENT: CHALLENGES, METHODS, AND AI-DRIVEN APPROACHES

Marcel LINEK, Petra BOBÍKOVÁ, Pavol PETROVIČ, Roman RIGÓ, Marek BELIČKA

Abstract

The digital transformation of waste management systems—encompassing Radio Frequency Identification (RFID)-equipped collection bins, Global Positioning System (GPS)-tracked vehicles, sensor-based smart containers, and pay-as-you-throw (PAYT) billing—generates continuous streams of data that carry significant privacy implications. These data intersect personal, commercial, and operational interests of municipalities, citizens, waste-collection companies, producer responsibility organisations (PROs), regulators, and researchers. Waste management presents sector-specific challenges that remain underexplored: the physical tangibility of waste enables re-identification even after digital anonymization; PAYT systems expose household behavioural patterns; and industrial waste data constitute commercially protected trade secrets. This paper analyses the anonymization requirements of all major stakeholder groups, reviews legal obligations, and proposes a layered anonymization workflow grounded in operational practice at waste-collection operators. Both classical and emerging AI-driven techniques are evaluated—including differential privacy, NLP/NER-based document redaction, OCR pipelines, computer-vision anonymization, Vision-Language Models (VLMs), agentic LLM pipelines, synthetic data generation, and federated learning. A strict requirement for all AI-driven methods is on-premise deployment. A comparative analysis against water, energy, and telecommunications utilities highlights what makes waste data unique.

Key words:
data anonymization; waste management; PAYT; differential privacy; privacy by design

JEL Classification Q53, K32, O33 https://doi.org/10.52665/ser20260106

INTRODUCTION

Waste management is undergoing rapid digitalization. Modern systems rely on RFID chips embedded in collection bins, GPS tracking of vehicles, fill-level sensors in smart containers, and weight-based PAYT billing to optimize collection logistics and enforce producer-responsibility obligations (Brighente et al., 2023). These technologies generate fine-grained data streams that are operationally indispensable but simultaneously encode sensitive information about individuals, businesses, and public infrastructure.

The privacy implications are broader than commonly assumed. A PAYT record links a specific bin chip to a household and, when combined with timestamps, reveals daily routines, holiday absences, dietary habits, and health indicators through pharmaceutical packaging (Brighente et al., 2023; Sampaio et al., 2023). Industrial waste data disclose production volumes and supply-chain relationships that companies classify as trade secrets. GPS logs of collection vehicles expose the operational topology of private businesses.

Despite this, waste management remains an underexplored domain in the data-privacy literature. Existing work either treats it as a peripheral use case of broader smart-city frameworks (Sampaio et al., 2023; Liu et al., 2017) or addresses IoT security threats at a hardware level (Brighente et al., 2023). A systematic, stakeholder-grounded treatment of anonymization requirements, applicable legal obligations, and AI-driven technical solutions in the waste sector is absent.

This paper fills that gap with the following contributions: (i) a structured analysis of anonymization needs for all major stakeholder groups; (ii) a review of legal obligations; (iii) an evaluation of classical and AI-driven anonymization techniques calibrated to waste-management data types; (iv) a five-layer

anonymization workflow derived from operational practice at Slovak waste-collection operators; and (v) a comparative analysis against water, heat, and telecommunications utilities.

2. Data in Waste Management and Privacy Challenges

2.1 Types of Collected Data

Waste-management information systems handle three categories of data, each with a distinct privacy profile (Brighente et al., 2023; Sampaio et al., 2023).

Static infrastructure data describe the physical collection network: container locations, bin capacities, and cadastral ownership records. Although relatively low-risk in isolation, static data enable household-level inference when combined with public registers.

Dynamic operational data are generated afresh for each collection event: RFID chip reads, container weights, GPS coordinates and routes of collection vehicles, precise timestamps, and fill-level sensor readings. A sequence of RFID reads with precise timestamps is sufficient to uniquely identify a household and reconstruct its daily routine, occupancy pattern, and consumption habits.

Producer data link waste quantities to their origin. For citizens, this means PAYT contracts and billing records. For businesses it means waste declarations and reports and carrier documents that can reveal production volumes and supply-chain relationships. In Slovakia, the split between industrial and municipal waste is approximately 80% to 20% by mass—industry produces roughly 10 million tonnes per year compared to 2.6–2.7 million tonnes of municipal waste—yet it is municipal data that create the greatest privacy challenges (Slovak Environmental Agency, 2023).

2.2 Stakeholder Anonymization Requirements

Data anonymization in waste management affects waste record-keeping, inspection, infrastructure planning, and public oversight. Each stakeholder has different motivations: municipalities and collection companies need data granularity; the state and analysts need aggregation; citizens and businesses protect privacy; technology providers implement technical solutions; and PROs require data accuracy to meet their legal obligations. Table 1 summarises the key tensions.

Table 1: Stakeholder Anonymization Requirements in Waste Management

Stakeholder	Role	Anonymization Need	Key Conflict
Municipality / City	System organizer, planner	Household-level data for planning; anonymized for public reports	Accuracy vs. citizen privacy
Citizen	Primary waste producer (PAYT payer)	Bin chip ID linked to billing only; no lifestyle profiling; right to be forgotten	PAYT enforcement vs. behavioural exposure
Business	Industrial/commercial producer	Production volumes hidden as trade secrets; ESG reporting preserved	Regulatory transparency vs. competitive exposure
Collection company	Operational collector, RFID records	Full detail for invoicing; anonymized before sharing	Liability reduction vs. reporting obligations
PRO (OZV)	Producer-responsibility fulfilment, audit	Pseudonymized material flows for audit; market shares hidden	Audit accuracy vs. B2B confidentiality
State / Regulator	Legislation, Eurostat reporting	Municipal-level aggregates;	Statistical quality vs. GDPR minimization
Inspection authority	Compliance verification, GDPR oversight	Pseudonymization with audit trail; DPIA for new tech	Control access vs. lawful processing
Technology provider	ERP, sensors, smart bins	Privacy-by-design at edge; PII separated from operational data	Scalability vs. bespoke municipal standards

Waste-collection companies operate at the intersection of all stakeholder interests: they must maintain identifiable records for invoicing (RFID chip → customer contract), yet share anonymized statistics with municipalities and PROs. Employee telemetry from vehicle tracking systems adds a further dimension: telematics data must be anonymized for general fleet analytics to avoid conflicts with labour law (EDPB, 2020).

PROs – Producer Responsibility Organisations (in Slovakia termed OZV) occupy a structurally unique position: their auditing function requires access to collection records, but they do not want to become controllers of citizen personal data. Anonymization at the household level lets them verify that a collection route was completed without ever knowing which household produced what volume. Conversely, their B2B data must remain confidential to prevent competitors from inferring market shares (Sampaio et al., 2023).

3. LEGAL AND REGULATORY FRAMEWORK

Anonymization is frequently discussed in the context of the General Data Protection Regulation (Regulation (EU) 2016/679) (GDPR), where it functions as a key mechanism for excluding data from the scope of personal data protection law. As clarified in Recital 26 GDPR, data that have been rendered truly anonymous are no longer considered personal data, whereas pseudonymized data remain subject to the full set of regulatory obligations. This distinction provides a central legal incentive for the implementation of anonymization techniques in data-driven systems.

However, in the context of waste management, a strictly GDPR-centric understanding of anonymization is inherently limited. Waste management systems generate and process heterogeneous datasets that, from a legal perspective, extend beyond personal data and may simultaneously encompass multiple categories of legally relevant information. These include, in particular, (i) personal data related to individuals (e.g., household waste production patterns), (ii) commercially sensitive information concerning business entities (e.g., industrial waste composition or volume), and (iii) operational or infrastructural data (e.g., collection routes or facility capacities).

From a legal perspective, anonymization should therefore be understood as a cross-cutting risk mitigation mechanism aimed at preventing the identification of legally protected interests, rather than merely the identification of natural persons. While GDPR focuses on the identifiability of individuals, other legal regimes—such as trade secret protection, contractual confidentiality obligations, or sector-specific regulation—may apply even where data no longer qualifies as personal data.

This broader understanding reveals a potential misalignment between technical and legal approaches to anonymization. Technical implementations often define anonymization in terms of the removal or transformation of direct identifiers, or through aggregation techniques. However, such measures may be insufficient from a legal standpoint if the resulting dataset still enables the inference of commercially sensitive information or reveals strategic or operational characteristics of specific entities.

Consequently, anonymization in waste management should be assessed in a multi-layered manner, taking into account not only the risk of re-identification of individuals, but also the potential for indirect identification or inference affecting corporate actors and critical infrastructure.

In addition, PAYT systems, which enable the identification of individual households, introduce specific privacy and societal risks. Such systems may facilitate detailed lifestyle profiling: patterns in pharmaceutical packaging may reveal health conditions; consumption habits may be inferred from waste composition; and temporal disposal patterns may disclose occupancy rhythms, including periods of absence. Furthermore, residents of lower-income neighborhoods may face an increased risk of stigmatization where waste-related performance indicators, such as recycling rates, are published at a granular, street-level resolution without adequate aggregation.

Accordingly, the adequacy of anonymization measures in this context must be assessed not only against the risk of individual re-identification, but also in light of their potential to enable inference,

profiling, and socially detrimental outcomes, thereby requiring evaluation across multiple legal and regulatory frameworks rather than exclusively under GDPR criteria.

4. ANONYMIZATION METHODS FOR WASTE MANAGEMENT DATA

Anonymization techniques fall into classical statistical methods applicable to structured tabular records, and AI-driven approaches that extend anonymization to unstructured documents, spatial data, and video streams. Both categories are required in waste management; their combination within a layered architecture is the recommended approach.

4.1 Classical Statistical Methods

Generalization and suppression reduce the precision of attribute values (e.g., exact address → postal-code zone) or remove them entirely. Applied to GPS coordinates, generalization to street-segment level eliminates household-level identifiability while preserving spatial density information needed for route planning.

Pseudonymization and tokenization replace real identifiers with cryptographically generated tokens. In waste-collection practice, the bin chip ID and customer contract number are the primary identifiers: at the moment of collection, the vehicle system replaces the real chip ID with an HMAC-SHA256 token using a daily-rotating key. Only the billing department retains the mapping table; all other system modules (dispatching, analytics, reporting) operate exclusively on tokens. This is operationally non-negotiable—collection operators cannot function without linking collections to customers for invoicing—but full anonymization of this link is possible once the billing purpose is fulfilled (Brighente et al., 2023).

k-Anonymity (Sweeney, 2002) guarantees that each published record is indistinguishable from at least $k-1$ others on quasi-identifier attributes. Applied to street-level waste statistics, a rule of $k \geq 5$ means data for a street with fewer than five households are merged with an adjacent segment before publication. This directly implements the operational requirement of collection operators who, when reporting to municipalities or PROs, must ensure that aggregated tonnage figures cannot be traced to individual addresses. k-Anonymity alone is vulnerable to homogeneity attacks when sensitive attribute values within a group are uniform; **l-diversity** and **t-closeness** extensions address this limitation.

Temporal generalization (time masking) addresses the fact that a precise collection timestamp exposes a household's daily routine. Rounding timestamps to one-hour slots is sufficient for route efficiency statistics while preventing behavioural inference.

4.2 AI-Driven Anonymization Methods

Classical methods are insufficient for unstructured data sources prevalent in waste management: scanned carrier documents, free-text ERP fields, and video streams from vehicle-mounted cameras.

To illustrate: a scanned delivery note accompanying a waste shipment may contain the driver's name, the client's company name, and a GPS-tagged pickup address. A human clerk would redact these manually; the AI pipeline described below performs the same task automatically, at scale, and without human access to the underlying data.

Differential Privacy (DP) (Dwork et al., 2006) provides a mathematically rigorous guarantee by injecting calibrated Laplace noise into query results. In waste management, DP is applied to aggregate reporting: recycling-rate statistics exported to municipalities or open-data portals are protected with $\varepsilon = 0.5$ noise, preserving aggregate accuracy while preventing individual inference. The same Laplace mechanism provides practical guarantees for location and mobility data (Duchi et al., 2013), making it directly applicable to GPS trajectory records from collection vehicles.

NLP/NER-based document redaction uses pre-trained transformer models (BERT, RoBERTa) fine-tuned for Named Entity Recognition to automatically detect and redact personal spans—person names, company names, addresses, VAT numbers, and GPS coordinates embedded in text—replacing them with category placeholders such as [PERSON] or [ORG] (Kaur et al., 2025). Redaction can replace identified

spans with category labels or synthetic substitutes generated by language models, preserving document readability for downstream analysis. This technique directly automates the manual export process currently used by collection operators when preparing data for commercial presentations or research handovers.

OCR integration enables retroactive anonymization of scanned paper documents (weighbridge slips, carrier documents, historical waste declarations) via an OCR-to-NER pipeline: Tesseract or Azure Document Intelligence converts scanned images to machine-readable text, a ByT5-based spell-correction module reduces OCR noise, and the NER redactor processes the cleaned text. Quality of anonymization depends critically on OCR accuracy; post-correction is therefore recommended before NER processing (Kaur et al., 2025; Brighente et al., 2023).

Computer vision anonymizes video streams from vehicle-mounted cameras in real time. A YOLOv8-based detection pipeline blurs faces and license plates before frames are encoded and transmitted, implementing Privacy by Design at the edge (Zhang et al., 2021; Angus and Duan, 2022). For applications requiring higher visual fidelity, GAN-based face replacement (CIAGAN; Maximov et al., 2020) preserves scene utility while eliminating biometric identifiability.

Spatial fuzzing applies a Laplace-distributed coordinate offset of 20–50 m to GPS points before publication, preserving spatial density patterns for route optimization while preventing point-level re-identification. Two complementary techniques are applied: (a) *aggregation to polygons*—precise coordinates are replaced by street-segment or neighborhood-zone codes in all external reporting; (b) *coordinate fuzzing*—a random Laplace offset is added before publication. Collection operators already practice a form of this by publishing route data aggregated to street level rather than individual stop coordinates (Brighente et al., 2023; Sampaio et al., 2023).

Vision-Language Models (VLMs) such as LLaVA and Llama 3.2 Vision advance document anonymization by processing scanned document images end-to-end, simultaneously interpreting layout and content without a separate OCR stage. This context-awareness is critical for complex waste-sector forms where positional semantics determine sensitivity. **A strict requirement is that VLMs are deployed exclusively as locally-hosted models** (e.g., via Ollama[1] or llama.cpp[2]): sending waste documents to cloud-based APIs constitutes a transfer of personal data under GDPR Article 28 and, where the provider is outside the EEA, a cross-border transfer under Chapter V.

Agentic LLM pipelines combine a locally-hosted LLM (e.g., Mistral 7B, Qwen2.5) with specialist tools—NER redactor, k-anonymity checker, spatial aggregator, audit logger—orchestrated via a ReAct (Reason-Act-Observe) loop. Given a document or dataset, the agent autonomously: (1) classifies the data type and applicable sensitivity level; (2) selects and invokes the appropriate anonymization tool; (3) verifies the output against re-identification criteria; and (4) iterates if residual risk is detected, logging each step for the GDPR audit trail. This replaces the labour-intensive manual export process used by collection operators. The entire pipeline runs on-premise and is auditable by the Data Protection Authority without exposing data to external parties.

Synthetic data generation produces statistically equivalent artificial datasets containing no real individuals. CTGAN and TVAE generative models trained on historical waste data—weights, timestamps, GPS zones—learn the underlying distributions and inter-variable correlations and sample new records that preserve aggregate statistics without retaining any original row (Bertino and Sandhu, 2005). For waste management, synthetic datasets enable municipalities and researchers to train route-optimization or fill-level prediction models without access to citizen PAYT data, and allow operators to demonstrate system capabilities in commercial presentations without exposing real customer or geographic information. Properly generated synthetic data carry no GDPR obligations, as they do not qualify as personal data.

[1] https://ollama.com

[2] https://github.com/ggml-org/llama.cpp

Federated learning (FL) enables collaborative AI model training across multiple municipalities without any raw data leaving the local environment. Each participant trains a local model update on their own dataset; only the model gradients (not the data) are aggregated by a central coordinator to produce an improved global model. For waste management, FL allows a shared fill-level-prediction or collection-route model to be trained across multiple Slovak municipalities without any city exchanging citizen-level PAYT records. When combined with differential privacy applied to the gradients (DP-SGD), FL provides a mathematically bounded privacy guarantee even against a malicious aggregator. The coordinator server never holds raw data, eliminating the primary GDPR data-controller liability from the aggregation step.

Note: Methods based on AI model inference (VLMs, agentic pipelines) must be deployed exclusively on local infrastructure. Cloud-hosted model APIs introduce GDPR transfer risk and audit opacity, making them incompatible with Privacy by Design (Art. 25 GDPR).

4.3 Method-to-Data-Type Mapping

Table 2 maps the waste-management data types to recommended anonymization methods, balancing privacy risk against operational constraints.

Table 2: Recommended Anonymization Methods by Waste-Management Data Type

Data Type	Privacy Risk	Recommended Method(s)	Deployment Constraint
Bin chip ID (RFID)	High – household link	Tokenization (HMAC-SHA256, daily key)	Billing module retains mapping; all others use tokens only
GPS collection point	High – re-id via cadastre	Spatial aggregation (street/zone) + Laplace fuzzing (20–50 m)	Route optimization requires density, not point precision
Collection timestamp	Medium – daily routine	Time-slot generalization (1-hour bins)	Schedule analytics need slot-level precision
Waste weight per bin	Medium – PAYT link	k-Anonymity ($k \geq 5$) per street segment; DP noise for open data	Internal invoicing requires exact weight
Carrier / declaration doc.	High – names, addresses, VAT	OCR → NER (BERT) or VLM end-to-end (LLaVA, local)	All inference on-premise; audit log required
Vehicle video stream	High – faces, licence plates	YOLOv8 edge detection + Gaussian blur / CIAGAN	Blurring on-vehicle before transmission
ERP free-text fields	Variable	NER fine-tuned; agentic LLM for context-aware redaction	Local LLM only; replaces manual export
Historical tabular records	Medium – aggregate re-id	Synthetic data (CTGAN / TVAE) for research / demo handover	Generator trained on-premise; output freely shareable
Cross-municipal AI training	High – citizen PAYT	Federated Learning + DP-SGD on gradients	Raw data never leave local node

5. CASE STUDY: ANONYMIZATION WORKFLOW FOR WASTE-COLLECTION OPERATORS

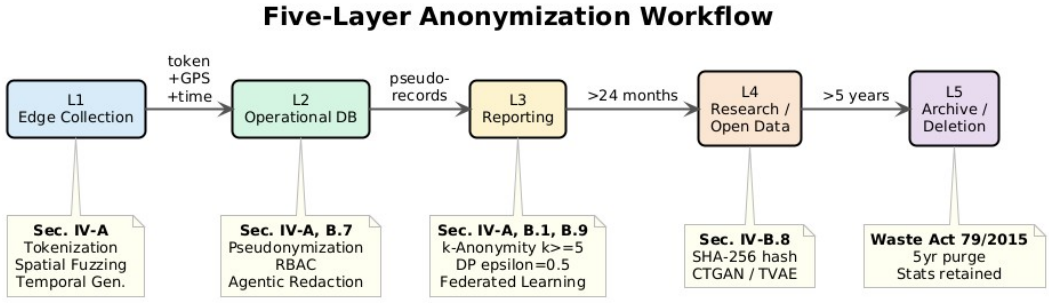
Waste-collection companies are simultaneously data controllers, data processors (acting on behalf of municipalities), and data subjects (employee telemetry). This section demonstrates how the methods described in Section 4 are composed into an integrated, five-layer workflow derived from operational

requirements at Slovak collection operators including Marius Pedersen, BRANTNER Slovakia, OLO, KOSIT, FCC Slovensko, and AVE SK, which together cover the majority of the Slovak municipal waste-collection market (Slovak Environmental Agency, 2023).

5.1 Five-Layer Anonymization Architecture

The architecture separates data by purpose and applies progressively stronger anonymization as data move away from their operational origin. Fig. 1 shows the data flow; each layer references the Section 4 methods it employs.

Figure 1: Five-Layer Anonymization Architecture for Waste-Collection Operators.



The five layers are as follows:

Layer 1 – Edge Collection (Sec. 4.1: Tokenization, Spatial Fuzzing, Temporal Gen.): The RFID reader on the vehicle replaces each bin chip ID with an HMAC-SHA256 token (daily-rotating key) before any data leave the vehicle. GPS coordinates are fuzzed with a Laplace offset of 20–50 m and timestamps are rounded to 5-minute slots. Raw identifiers never transit the network.

Layer 2 – Operational Database (Sec. 4.1, 4.2: Pseudonymization, RBAC, Agentic Redaction): The central ERP stores pseudonymized records; only billing staff hold the token-to-customer mapping. Incoming scanned carrier documents are routed through the on-premise agentic anonymization pipeline (Sec. 4.2) before storage, replacing manual redaction.

Layer 3 – Reporting to Municipalities and PROs (Sec. 4.1, 4.2: k-Anonymity, DP, FL): Records are aggregated to street-segment level with $k \geq 5$. Differential privacy noise ($\varepsilon = 0.5$, Laplace mechanism) is added to weight totals. For cross-municipal model training, federated learning with DP-SGD is applied so that raw records never leave the operator's server.

Layer 4 – Research and Open Data (Sec. 4.2: Full Anonymization, Synthetic Data): Records older than 24 months are fully anonymized: tokens replaced with non-reversible SHA-256 hashes, temporal precision reduced to month-year, spatial precision to neighborhood level. For research datasets, CTGAN/TVAE-based synthetic data generation produces GDPR-free datasets preserving aggregate distributions.

Layer 5 – Archive and Deletion (Waste Act No. 79/2015): Records older than the statutory five-year retention period are automatically purged. Aggregated statistical records (no personal data) are retained indefinitely, preserving institutional memory of waste flows.

This architecture operationalizes GDPR Articles 5, 17, 25, and 35. In the event of a data breach, anonymized tokens at Layers 2–4 carry no notification obligation, transforming compliance risk into a manageable technical property (Brighente et al., 2023; Sampaio et al., 2023).

5.2 Vehicle Camera Pipeline

This pipeline instantiates the computer-vision methods of Section 4.2 in an operational on-vehicle context. YOLOv8-nano detects faces and license plates at ≈25 fps; a 31×31 Gaussian kernel blurs detected regions before frame encoding. The anonymized frame—not the raw video—reaches the central server. For contamination-detection applications, the CIAGAN face-replacement method (Maximov et al., 2020) is

applied instead of blurring. VLM-based post-processing of selected frames is executed on-premise via a locally-hosted model.

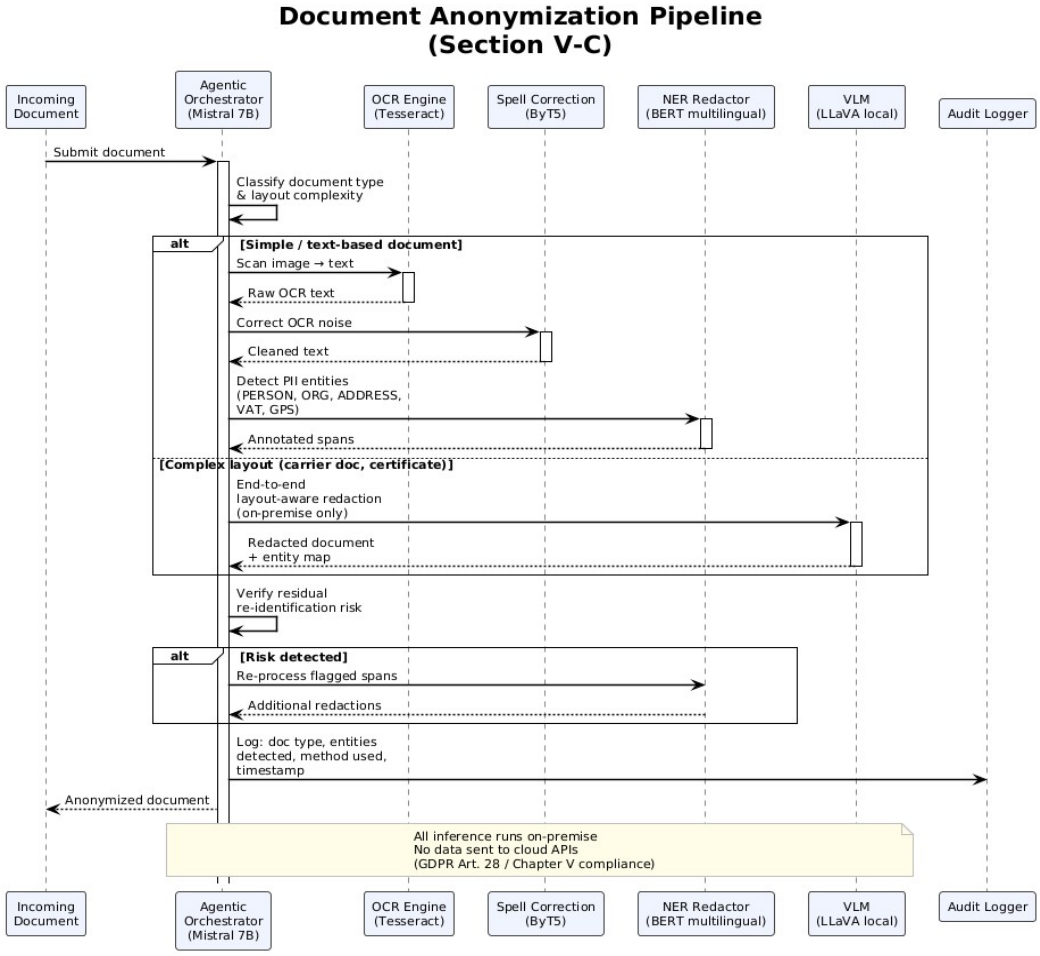
In practice, the pipeline runs on an embedded GPU module (e.g., NVIDIA Jetson Orin) mounted in the collection vehicle. Processing latency stays below 40 ms per frame, ensuring real-time operation without buffering raw footage. Anonymized frames are stored locally for 24 hours for operational review and then purged automatically, in line with the data-minimisation principle of Article 5(1)(c) GDPR.

5.3 Document Anonymization Pipeline

In simple terms: every document entering the system is automatically checked for personal data and sensitive identifiers, which are replaced by neutral placeholders before the document is stored or shared.

Fig. 2 shows the document anonymization sequence. Legacy scanned documents follow the OCR→NER path (Tesseract + ByT5 spell correction + BERT-base-multilingual NER). New documents with complex layouts (multi-party carrier forms, weighbridge certificates) are processed end-to-end by a locally-hosted VLM (LLaVA). Both paths are orchestrated by the agentic pipeline, which classifies incoming documents, selects the appropriate path, verifies residual re-identification risk, and logs each decision for the GDPR audit trail.

Figure 2: Document Anonymization Pipeline: OCR/NER path for legacy documents and VLM end-to-end path for complex layouts, orchestrated by an on-premise agentic system.



6. Comparison with Other Regulated Utility Sectors

Waste management shares the regulated-utility context with water supply, district heating, and telecommunications. Table 3 summarises the key dimensions of comparison.

Table 3: Anonymization Comparison Across Regulated Utility Sectors

Dimension	Waste	Water (URSO)	Heat / Energy	Telecommunications
Data cadence	Episodic (collection events, weekly)	Continuous (smart meter, real-time)	Continuous (smart meter, seasonal)	Real-time (CDR, location, browsing)
What data reveal	Lifestyle, health indicators, occupancy	Occupancy, health (nightly water draws)	Economic vulnerability, occupancy	Location, social graph, interests
Re-id risk	Medium – RFID + cadastre + timestamp	High – flow curves near-biometric	High – seasonal correlation	Very high – mobility traces
Primary method	Tokenization, k-anonymity, DP, spatial fuzzing	Flow-curve perturbation, DP	Zone-level aggregation	DP (trajectory), CDR aggregation
Regulation	GDPR + Waste Act + WFD	GDPR + Water Act + URSO	GDPR + Energy Act	GDPR + ePrivacy + NIS2
Unique challenge	Physical dumpster diving bypasses digital anonymization	Leak detection needs sub-hourly precision	Seasonal correlation degrades anonymization	Re-id from mobility even after DP

The key differentiator of waste data is its physical tangibility: digital anonymization of database records does not prevent inference from direct inspection of physical waste (dumpster-diving attacks). This means digital anonymization must be complemented by physical-security controls—a challenge with no parallel in digital utility sectors (Brighente et al., 2023; Sampaio et al., 2023).

Water utilities (URSO-regulated in Slovakia) face the most analogous challenge: high-frequency smart-meter data are effectively a biometric signature of a household. Unlike waste, water consumption is continuous—a single day's water-draw curve reveals the number of occupants, nightly health conditions (nocturnal polyuria), and when the apartment is empty. Anonymization must therefore protect flow curves, not just totals, while preserving sufficient precision for network-leak detection. The recommended approach—trusted execution environments running leak-detection algorithms on non-exported raw data, with only aggregated daily totals published—is directly applicable to future high-frequency PAYT systems that weigh bins at every emptying (Sampaio et al., 2023; Brighente et al., 2023).

Heat and energy suppliers present a different challenge: low heating consumption in winter can signal energy poverty or an empty apartment (a burglary risk). Anonymization must suppress socio-economic inference while satisfying energy-efficiency regulations. The seasonal correlation of heat data with weather patterns also complicates anonymization: residualizing out weather effects can re-expose individual consumption anomalies (Sampaio et al., 2023; Liu et al., 2017).

Telecommunications operators represent the most complex anonymization challenge. Mobility traces and call-detail records (CDRs) enable re-identification even after differential privacy is applied, because trajectory data contain unique spatio-temporal fingerprints (Liu et al., 2017). The key difference from waste data is *dynamic trajectories vs. static points*: waste anonymization deals primarily with fixed container locations and scheduled collection events, making it technically far more tractable than telco. The unique challenge with no parallel in digital utility sectors is physical dumpster diving, which requires physical-security controls as a necessary complement to digital anonymization (Brighente et al., 2023; Liu et al., 2017; EDPB, 2020).

CONCLUSION

This paper has presented a systematic, stakeholder-grounded analysis of data anonymization in waste management. Three principal findings emerge.

First, waste-management data are more privacy-sensitive than the sector traditionally acknowledges. The combination of RFID identifiers, GPS coordinates, and precise timestamps uniquely identifies households and exposes behavioural, health, and economic information. Industrial waste data constitute trade secrets requiring protection under commercial law in addition to GDPR. The multi-stakeholder structure creates irreconcilable tensions that no single technique resolves.

Second, AI-driven methods substantially extend classical anonymization to the unstructured data types prevalent in waste management. NLP/NER pipelines with OCR integration automate manual document-redaction. Computer-vision blurring and GAN-based face replacement enable Privacy by Design compliance for vehicle camera systems. VLMs enable context-aware end-to-end document anonymization; agentic LLM pipelines orchestrate the full workflow with a logged audit trail; synthetic data generation produces GDPR-free datasets; and federated learning enables cross-municipal model training without raw-data sharing. Together, these techniques address all nine data types identified in the operational taxonomy (Table 2). A critical deployment requirement is that all AI inference runs on local infrastructure—not cloud APIs.

Third, the five-layer workflow—edge pseudonymization, role-controlled operational database, k-anonymous reporting, DP-protected open data, and scheduled archive purge—operationalizes GDPR Articles 5, 17, 25, and 35 in a sector-specific manner. Its most important practical benefit is liability reduction: anonymized tokens released in a breach carry no GDPR notification obligation, transforming risky data into safe business assets without sacrificing analytical value.

From a practical standpoint, several challenges remain. (1) The diversity of legacy document formats in waste management makes it difficult to deploy a single NER model with consistently high recall across all operators. (2) The absence of labelled waste-sector datasets limits the benchmarking of anonymization quality. (3) Physical re-identification through direct waste inspection has no technical countermeasure — it requires organisational and legal controls that fall outside the scope of data anonymization alone.

Compared to water and energy utilities, waste management benefits from episodic (rather than continuous) data cadence, which simplifies anonymization. Compared to telecommunications, waste deals with fixed points rather than dynamic trajectories, making re-identification technically more tractable. As noted in Section 6, the absence of a purely digital countermeasure to physical waste inspection remains the sector-specific open problem.

Future work will focus on: quantifying re-identification risk on real Slovak waste-management datasets using membership-inference attacks; developing a labelled waste-sector NER corpus for benchmarking document-redaction models; and evaluating federated learning for cross-municipal model training without raw-data sharing.

ACKNOWLEDGMENTS

Funded by the EU NextGenerationEU through the recovery and resilience plan for Slovakia under the project no. 09105-03-V02-00056: AI4WasteManagement: Research on prediction and optimization of waste management processes

REFERENCES

Angus, A. and Duan, Z. (2022). Real-time video anonymization in smart city intersections. In *Proceedings of the IEEE MASS Conference*.

Bertino, E. and Sandhu, R. (2005). Database security—concepts, approaches, and challenges. *IEEE Transactions on Dependable and Secure Computing*, 2(1), 2–19. https://doi.org/10.1109/TDSC.2005.9

Brighente, A., Conti, M., Di Renzone, G. and Peruzzi, G. (2023). Security and privacy of smart waste management systems: A cyber-physical system perspective. *IEEE Internet of Things Journal*, 10(10), 8580–8592. https://doi.org/10.1109/JIOT.2023.3322532

Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In *Proceedings of the IEEE FOCS*, 429–438. https://doi.org/10.1109/FOCS.2013.53

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of TCC*, LNCS vol. 3876, 265–284. https://doi.org/10.1007/11681878_14

EDPB (2020). *Guidelines 01/2020 on Processing Personal Data in the Context of Connected Vehicles and Mobility Related Applications*. European Data Protection Board, Brussels.

Kaur, R. et al. (2025). A survey on privacy preservation techniques in IoT systems. *Sensors*, 25(22), 6967. https://doi.org/10.3390/s25226967

Liu, X., Heller, A. and Nielsen, P. S. (2017). CITIESData: A smart city data management framework. *Knowledge and Information Systems*, 53(3), 699–722. https://doi.org/10.1007/s10115-017-1051-3

Maximov, M., Elezi, I. and Leal-Taixé, L. (2020). CIAGAN: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE CVPR*. arXiv:2005.09544

Muştu, M. İ. and Ekenel, H. K. (2025). Assessing the use of face swapping methods as face anonymizers in videos. *arXiv preprint arXiv:2501.04955*.

Ribačič, S. and Fratrič, I. (2016). De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47, 131–151. https://doi.org/10.1016/j.image.2016.05.020

Sampaio, S. et al. (2023). Collecting, processing and secondary using personal and (pseudo)anonymized data in smart cities. *Applied Sciences*, 13(6), 3830. https://doi.org/10.3390/app13063830

Slovak Environmental Agency (SŽP) (2023). *Report on the State of the Environment of the Slovak Republic*. Ministry of the Environment of the Slovak Republic, Banská Bystrica.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. https://doi.org/10.1142/S0218488502001648

Zhang, Z., Cilloni, T., Walter, C. and Fleming, C. (2021). Multi-scale, class-generic, privacy-preserving video. *Electronics*, 10(10), 1172. https://doi.org/10.3390/electronics10101172

Contact

Marcel Linek, D2B s.r.o.
Bratislava, Slovakia
e-mail: marcel.linek@d2b.sk

Petra Bobíková, Asseco Central Europe, a. s.
Bratislava, Slovakia
e-mail: petra.bobikova@asseco-ce.com

Mgr. Pavol Petrovič, PhD., Asseco Central Europe, a.s.
Bratislava, Slovakia
e-mail: pavol.petrovic@asseco-ce.com

Roman Rigó, Asseco Central Europe, a. s.
Bratislava, Slovakia
e-mail: roman.rigo@asseco-ce.com

Marek Belička, Asseco Central Europe, a. s.
Bratislava, Slovakia
e-mail: marek.belicka@asseco-ce.com