Datasent

WHITEPAPER · COMPANION PAPER

A Trusted Setup for Bandwidth-Minimal, Lossless Data Tokenisation

Raw data never traverses the wire. Reconstruction is governed and exact.

AUTHORED BY
Datasent Research

CONTACT
info@datasent.com · datasent.com

EDITION
November 2026

CONTENTS

Table of contents

A condensed treatment of the trusted-setup transmission protocol. Full proofs, residual range bounds, floating-point determinism, and code-to-mathematics mapping are deferred to the appendices in the full paper.

·	Abstract	03
1	Introduction	03
2	System overview	04
3	Canonicalisation	05
4	Integer-lossless representation	06
5	Trusted setup model	07
6	Residual-only transmission	08
7	Threshold and custodian reconstruction	09
8	Integrity and verification	10
9	Relation to zero-knowledge proofs	10
10	Applications	11
11	Limitations	12
12	Closing	12

ABSTRACT

A Trusted Setup for Bandwidth-Minimal, Lossless Data Tokenisation

Companion paper to the Datasent research series. Datasent Research · info@datasent.com · datasent.com · 2026.

ABSTRACT

Modern data systems increasingly require sharing, processing, and analysing sensitive data across organisational and network boundaries. Traditional approaches transmit raw data or compressed archives, introducing bandwidth costs, privacy risk, and compliance complexity. We present a trusted setup–based protocol for lossless data tokenisation that enables exact reconstruction of original data while transmitting only minimal residual information — never the data itself. The approach combines deterministic canonicalisation, integer-lossless representation, and threshold reconstruction techniques inspired by cryptographic trusted setups. The result is a practical system in which raw data never traverses the wire, bandwidth is minimised, and reconstruction is gated by explicit authorisation — without requiring zero-knowledge proofs or heavy cryptography in the initial deployment.

SECTION 1

1 Introduction

Data gravity has become a dominant constraint in modern systems. Large volumes of sensitive data — text, logs, images, audio, video, and structured records — are expensive to move and risky to share. At the same time, downstream consumers (analytics platforms, ML pipelines, auditors, and AI systems) increasingly require access to these datasets for model training and evaluation, benchmarking and analytics, compliance and audit replay, and safe integration with LLM-based workflows.

Standard compression techniques reduce size, but they do not address:

- **Provenance and determinism.** Whether the representation is reproducible and stable across environments.
- **Fine-grained trust and authorisation.** Who can reconstruct, when, and under what policy.
- **Partial reconstruction workflows.** Enabling analytics and ML without always materialising raw data.

We propose a lossless tokenisation protocol with a trusted setup, in which raw inputs are transformed into canonical form, a deterministic representation is fit, only residuals (or shares thereof) are transmitted, and exact reconstruction is possible only under explicit trust conditions.

> **DESIGN GOAL**
>
> *Make "raw data movement" an exception rather than the default. Systems exchange compact, governed artifacts (tokens) suitable for downstream workflows while minimising bandwidth and risk.*

SECTION 2

2 System overview

2.1 The golden path

All data flows through the following invariant sequence:

The expanded sequence is: **raw input → canonicalisation → representation (lossless) → segmentation → tokenisation → residual / share transmission → threshold reconstruction (authorised)**. Deterministic canonicalisation and basis fitting precede lossless tokenisation. Only residuals and/or shares traverse boundaries; reconstruction is explicit and authorised.

2.2 Roles and parties

Three logical actors participate in the protocol.

ACTOR	ROLE
Data Holder (DH)	Possesses raw data X and its canonical form Y. Performs local fitting and transmits residuals.
Reconstruction Party (RP)	Does not possess X. Participates in reconstruction under authorisation.
Custodian (C)	Optional gatekeeper that authorises reconstruction and/or holds secret shares.

In a typical enterprise deployment the DH is a customer-controlled environment, the RP is an analytics or ML platform, and the Custodian is a governance service. These roles may map to different organisations, services, or trust domains.

SECTION 3

3 Canonicalisation

Lossless reconstruction requires that representation be deterministic and unambiguous. Canonicalisation ensures that equivalent inputs — two CSVs with the same records but different ordering, or text with different encodings — produce identical internal representations.

3.1 Canonical types

We define three canonical data types.

- `CanonicalMatrix`: integer matrix $Y \in \mathbb{Z}^{N \times D}$.
- `CanonicalBlob`: byte sequence $\mathcal{B} \in \{0,1\}^*$.
- `CanonicalBundle`: structured collection of matrices and blobs (e.g. video: bytes + motion + keyframes).

For lossless tokenisation we operate primarily on `CanonicalMatrix`. `CanonicalBlob` and `CanonicalBundle` remain important for provenance, modality-aware lossy representations, and semantic extraction.

3.2 Determinism guarantees

Canonicalisation enforces deterministic ordering (row-major for images, stable row ordering for tables), fixed integer quantisation for numeric inputs (fixed-point encoding for floats), and explicit metadata for scaling, layout, and segmentation hints. These guarantees are prerequisites for trusted reconstruction: the same canonical input must yield the same tokenisation decisions and the same decode behaviour across environments.

SECTION 4

4 Integer-lossless representation

Given a canonical integer matrix Y, we seek a representation

$$Y = \hat{Y} + R, \tag{1}$$

in which $\hat{Y}$ is a deterministic approximation derived from shared parameters, R is an integer residual (often small), and reconstruction is exact.

4.1 Polynomial / basis representation

We segment rows of Y into windows. For each segment s,

$$\hat{Y}_s = B_s C_s, \tag{2}$$

where $B_s \in \mathbb{R}^{L \times K}$ is a deterministic basis matrix and $C_s \in \mathbb{R}^{K \times D}$ are fitted coefficients. Residuals are

$$R_s = Y_s - \mathrm{round}(B_s C_s), \tag{3}$$

and stored as integers (int16 or int32) and compressed losslessly. Range selection between int16 and int32 is governed by the empirical bound $\|Y^{(s)} - \mathrm{round}(B^{(s)} C^{(s)})\|_\infty \leq M_s$ on each segment.

4.2 Losslessness guarantee

> **THEOREM · EXACTNESS** (Trusted Setup, A.2)
>
> *If B_s is deterministic, rounding rules are fixed, and residual bounds are respected, then*
>
> $$Y_s = \mathrm{round}(B_s C_s) + R_s$$
>
> *is exact. Under deterministic-decode arithmetic, $\hat{Y} = Y$ on every canonical matrix.*

This is the core of the lossless v3 codec. The implementation detail that makes it robust in practice — computing $R^{(s)}$ in int64 before downcasting after range checks — is given in the full paper.

SECTION 5

5 Trusted setup model

5.1 Definition

> **TRUSTED SETUP**
>
> *All parties agree on deterministic representation parameters **before** any data-dependent computation occurs. This includes basis type and parameters, segmentation strategy, and quantisation and rounding rules. This is **not** a cryptographic ceremony and does not require secrecy of setup parameters.*

5.2 What is trusted

The distinction between agreed and derived material is the structural core of the protocol.

TRUSTED · PUBLIC, AGREED	NOT TRUSTED · PRIVATE, DERIVED
Representation definition	Residual values
Basis generation rules	Coefficients derived from private data
Segment boundaries	

The trusted material describes the *class* of representations; the untrusted material is what makes any specific dataset reconstructable. The two are deliberately kept on separate channels.

SECTION 6

6 Residual-only transmission

6.1 Protocol steps

1. **Shared setup.** DH and RP agree on representation parameters (basis type, segmentation, rounding rules).
2. **Local fitting (DH).** DH computes C_s and R_s locally. Raw data Y never leaves the DH environment.
3. **Transmission.** DH sends only residuals R_s, or secret shares of C_s and/or R_s.
4. **Reconstruction (RP).** RP regenerates B_s locally from the agreed setup parameters and combines it with the received residuals to reconstruct Y_s.

> **INVARIANT**
>
> *Raw data Y is never transmitted. What crosses the network is only the residual — the unpredictable part.*

6.2 Bandwidth analysis

Let $|Y|$ be the raw data size, $|C|$ the coefficient size, and $|R|$ the residual size. For structured data, empirically,

$$|C| \ll |Y|, \qquad |R| \ll |Y|. \tag{4}$$

The transmitted payload is $|R|$ or $|C| + |R|$, often orders of magnitude smaller than $|Y|$. The bandwidth reduction is at the rate $|R|/|Y|$ in the residual-only configuration.

SECTION 7

7 Threshold and custodian reconstruction

7.1 2-of-2 model

In the simplest model, DH holds (or controls release of) C_s and RP holds (or receives) R_s. Reconstruction requires both. Neither party can recover Y_s unilaterally.

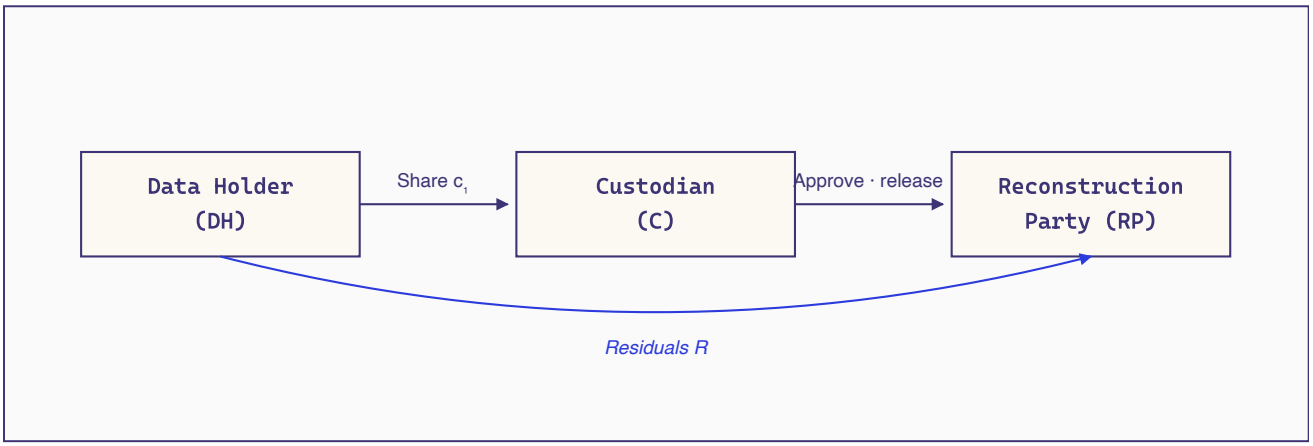

Figure 1. *Trusted setup with 2-of-2 shares. No single party can reconstruct alone; custodian gating adds explicit authorisation.*

7.2 k-of-n generalisation

Using Shamir secret sharing or XOR splitting, coefficients or residuals are split into n shares, and any k shares suffice to reconstruct. This enables multi-party governance, regulatory controls, and distributed trust.

7.3 Custodian gating

A custodian service may store one or more shares, require approval before releasing them, and log reconstruction events. Reconstruction becomes an explicit, auditable action rather than an implicit consequence of data possession.

SECTION 8

8 Integrity and verification

8.1 Deterministic verification

Because representation is deterministic, reconstructed data can be re-canonicalised and its hash compared to a stored digest. Tampering or corruption is detectable without requiring zero-knowledge proofs.

8.2 Checksums and hashes

Residual payloads include CRC32 or SHA-256, segment-level hashes, and optional prediction digests. The integrity chain covers:

$$\text{raw input} \rightarrow \text{canonical form} \rightarrow \text{tokenisation} \rightarrow \text{residual payload} \rightarrow \text{reconstruction}.$$

Each stage produces a hash that the next stage can verify, so end-to-end provenance is preserved even when only the residual crosses the wire.

SECTION 9

9 Relation to zero-knowledge proofs

This system does not require zero-knowledge (ZK) proofs to function. However, ZK can be layered to prove:

- Residuals correspond to a valid fitting.
- Coefficients satisfy error bounds.
- Reconstruction occurred under authorisation without revealing raw data.

ZK hooks are explicitly left as future work to avoid over-engineering the initial deployment. The architecture is designed to support this extension; the decoding computation is arithmetisable, and per-segment circuit size is $O(L_s \cdot K_s \cdot D)$.

SECTION 10

10 Applications

The same protocol serves four distinct application classes, with no modification to the encoding layer.

10.1 Call centres and financial data

Share payment propensity signals and audit-replay datasets without sharing transcripts or raw records broadly. The custodian gates reconstruction to authorised parties only.

10.2 Video analytics

Share motion tokens and residual summaries instead of raw footage. Reconstruct only under explicit policy. AI analytics operate on tokens; raw frames never leave the site.

10.3 Federated ML

Aggregate tokenised residual-based features across organisations while keeping raw data local. Each organisation holds its own residuals; a custodian authorises the reconstruction needed for model training.

10.4 LLM fine-tuning

Tokenise datasets into reconstructable, auditable artifacts for controlled fine-tuning workflows. Provenance is maintained from raw collection through to model training.

SECTION 11

11 Limitations

The protocol is deliberately scoped. The following are out of band for the initial deployment.

- Trusted setup assumes honest parameter agreement between parties.
- Representation choice may leak structural information (e.g. motion magnitude in video).
- The system is not designed to resist side-channel attacks.
- ZK guarantees are not yet implemented.

Each of these is independently addressable as the deployment matures. The architecture is structured to admit each extension without changes to the residual-transmission layer.

SECTION 12

12 Closing

We presented a trusted setup–based, lossless tokenisation protocol that minimises bandwidth, preserves privacy, and enables exact reconstruction under explicit authorisation. The system is deterministic, practical, implementable today, and extensible to cryptographic verification.

This approach reframes data sharing as **token exchange** rather than raw data movement. The central invariant of the protocol is

$$X \xrightarrow{\text{canon}} Y \xrightarrow{\text{fit}} (C, R) \xrightarrow{\text{wire: } R} \hat{Y} = Y, \tag{5}$$

in which the wire carries only the residual, and reconstruction is gated by explicit authorisation. Together with the foundational framework, this gives Datasent its operating principle: privacy and progress, held in balance, by construction.

For the full mathematical framework, see the foundational paper. For complete formal proofs, residual range bounds, floating-point determinism guarantees, and the code-to-mathematics mapping, see the appendices of the full A Trusted Setup for Bandwidth-Minimal, Lossless Data Tokenization paper. For commercial enquiries, contact ***info@datasent.com****.*