Clean Source Data

CLIENT GUIDE 4 OF 4

Protecting Personal Information

How personal information is removed or replaced inside your office before any data reaches an AI company, while keeping the business value intact.

PREPARED BY
Clean Source Data

DOCUMENT
Protecting Personal Information

EDITION
November 2026

CLEAN SOURCE DATA | CLIENT GUIDE 4 OF 4

Protecting Personal Information

A short guide to redaction, transformation, HIPAA Safe Harbor, and what is kept or left out of your data package.

01 Why This Matters

AI buyers want data that shows how real businesses work. They do not want, and should never receive, the identities of the people in it. Before anything is delivered, personal information is either **redacted** or **transformed**, on-site, inside your office. Knowing the difference helps you decide how your data is protected and what it is worth.

02 Redaction vs. Transformation

REDACTION

Removes sensitive details

Identifiers are taken out entirely.

```
"Maria called Dana at Oakline Dental"
becomes
"[NAME] called [NAME] at [COMPANY]"
```

Best for details that should never appear at all.

TRANSFORMATION

Replaces them with safe stand-ins

Identifiers are swapped for realistic substitutes, so relationships survive.

```
"Maria called Dana at Oakline Dental"
becomes
"Elena called Jordan at Company A"
```

Keeps who worked with whom, roles and context.

WHY TRANSFORMATION IS USUALLY WORTH MORE

Buyers value data that shows how people, roles and decisions connect. Transformation protects identities while keeping those connections, so the data stays useful and typically earns more than heavily redacted data.

03 Consistency Across Your Data

Good transformation is consistent across the whole package. The same real person or company always maps to the same substitute, and roles stay attached:

ORIGINAL	ALWAYS BECOMES
Maria Delgado	Elena Brooks
Oakline Dental	Company A
Dana, office manager	Jordan, office manager (role preserved)

The substitution key is created and kept on the offline device in your office. It is never delivered with the data and is destroyed after delivery. Every license also prohibits the buyer from attempting to re-identify anyone. (Names above are fictional examples.)

04 The HIPAA Safe Harbor Standard

We apply the HIPAA Safe Harbor method (45 CFR §164.514(b)(2)) to every client, in healthcare or not, because it is the strictest common standard. It requires removing 18 types of identifiers and having no actual knowledge that what remains could identify someone. Data that meets it is no longer protected health information.

1 Names	2 Geographic detail smaller than a state	3 All dates except the year
4 Phone numbers	5 Fax numbers	6 Email addresses
7 Social Security numbers	8 Medical record numbers	9 Health plan numbers
10 Account numbers	11 License and certificate numbers	12 Vehicle identifiers
13 Device identifiers and serial numbers	14 Web URLs	15 IP addresses
16 Biometric identifiers	17 Full-face photos and similar images	18 Any other unique identifying code

05 How Our Process Works

1. **On-site extraction.**
   Approved data is exported to an encrypted drive and loaded onto an offline device at your location. Nothing is sent over the internet.

2 **Automated identification.**
Software flags likely identifiers in text, files and image metadata, and proposes redactions or substitutions.

3 **Human review.**
Our team checks samples to confirm identities are protected and the business context is intact.

4 **Your approval and quality check.**
You review a sample and approve it in writing. The package is checked against the agreed criteria before it leaves.

5 **Buyer's independent review.**
After delivery, the AI company runs its own internal analysis to confirm the data is properly de-identified and no sensitive information remains. Anything it flags comes back to us to fix before acceptance.

TWO LAYERS, NOT ONE

Many brokers pass data to the buyer and let the buyer do the cleaning. With Clean Source Data, your data is de-identified inside your office first and then verified again by the buyer, so two separate teams check that nothing sensitive gets through.

06 What Is Kept, and What Is Left Out

USUALLY KEPT, THEN TRANSFORMED	USUALLY LEFT OUT ENTIRELY (NOT EXHAUSTIVE)
Names of people and companies	**Privileged legal material:** legal advice, litigation strategy, legal-hold material
Personal emails, phone numbers and addresses	**Credentials and secrets:** passwords, API keys, access tokens
Dates (reduced to the year where required)	**Financial account data:** payment card and bank account numbers
Health information, de-identified to Safe Harbor	**Biometric data**, and data about minors unless cleared case by case
Business operations information	**Third-party licensed content** you do not have the right to share

07 Your Controls

- You set the exclusion list before anything is prepared.
- You approve a sample of the clean data before anything is shared with any buyer.
- Every buyer independently verifies the de-identification before accepting the data.
- Rules on re-identification, retention and deletion are written into every license you approve.
- If a rights or privacy question comes up, we pause the process until it is resolved.

Questions?

Your Clean Source project lead is your single point of contact.

info@cleansourcedata.com
cleansourcedata.com