



Woodway Assurance

■ Trust in Data

Your Most Valuable Data is Trapped in Text

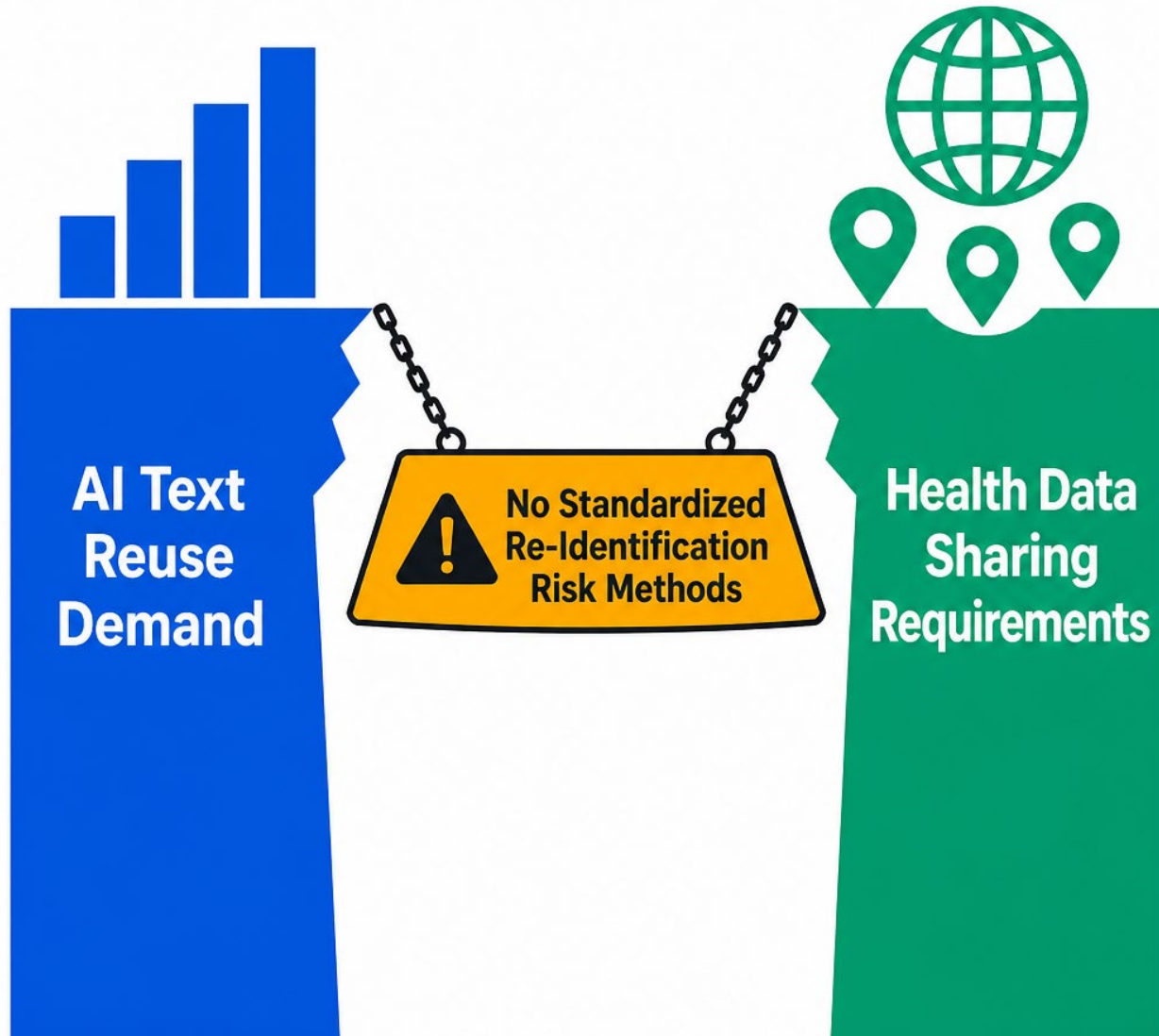
Khaled El Emam

kelemam@woodway-assurance.com

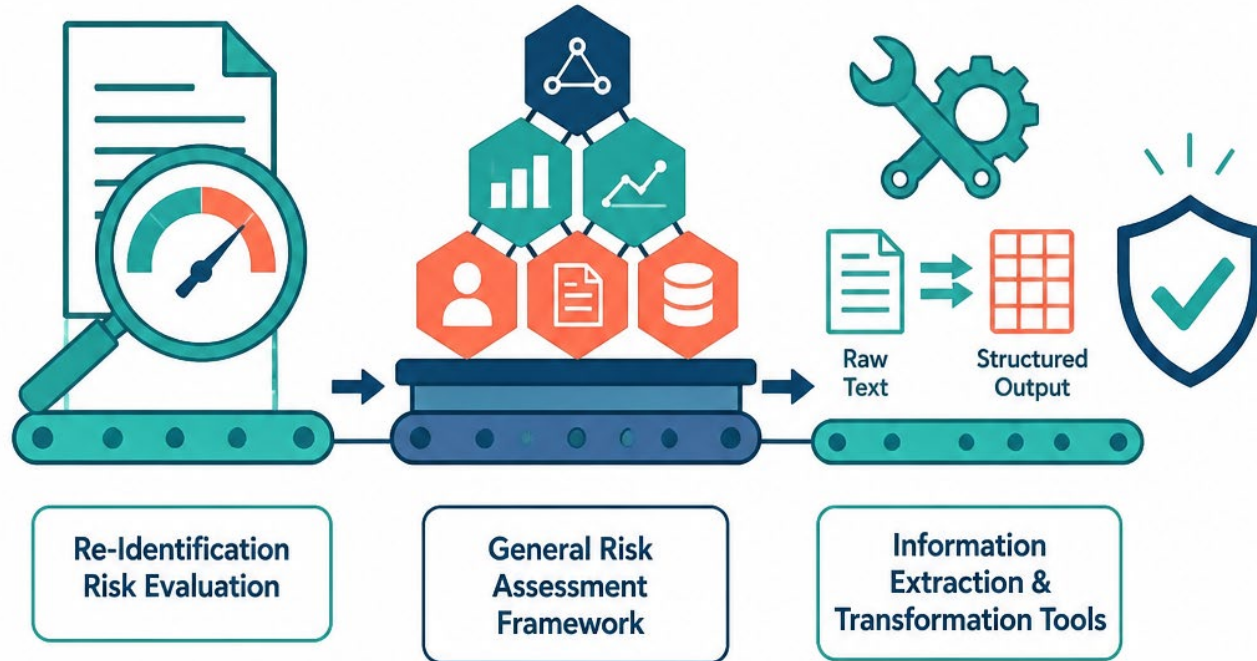
Jiacheng Hou

jhou@ehealthinformation.ca

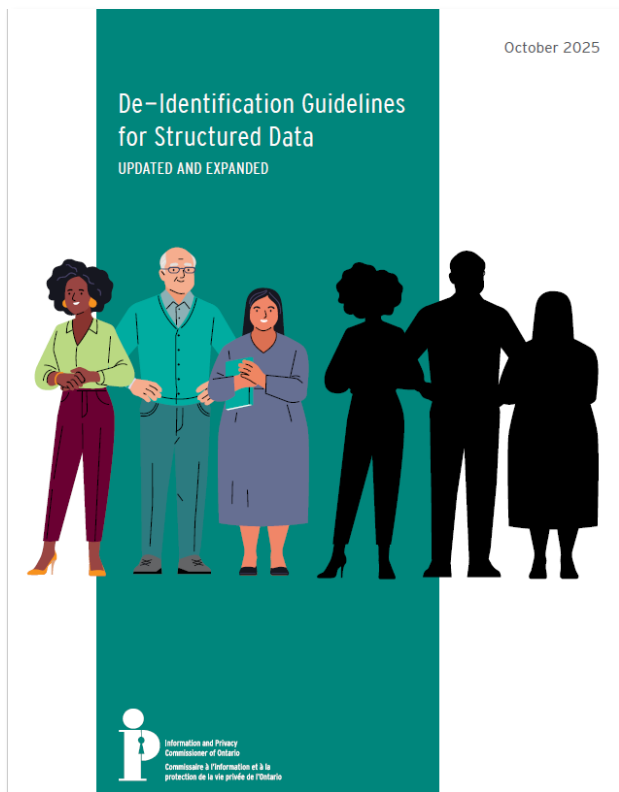
What We Are Seeing



How Are We Solving These Problems?



Ontario de-identification guidelines



- A recent set of operational guidelines from the Ontario privacy commissioner (2025 edition)
- Includes multiple checklists to allow organizations to have more certainty around what they need to do
- The process described enables organizations to balance data utility with privacy protection

Direct and Indirect Identifiers (examples)

DIRECT IDENTIFIERS

- Name
- Email address
- SIN / SSN
- Biometrics
- Health insurance number
- Full residential address

INDIRECT IDENTIFIERS

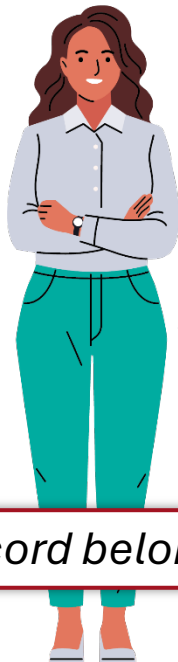
- Postal code / ZIP code
- Age / DoB
- Race / ethnicity / language
- Income
- Profession
- Number of children
- Marital status
- Visible characteristics (e.g., mobility devices)
- Dates of important events (e.g., marriage, death)

More examples of direct identifiers

- account numbers;
- certificate license numbers;
- vehicle identifiers and serial numbers, including license plates;
- device identifiers and serial numbers;
- URLs;
- IP address numbers;
- biometric identifiers (including finger and voice prints);
- names;
- street addresses (other than town, city, province and postal code);
- telephone numbers;
- fax numbers;
- e-mail addresses;
- Social Insurance Numbers;
- medical records numbers;
- health insurance numbers;
- full face photos

Identity disclosure is when a person's identity is assigned to a record

Sally



Indirect Identifiers

New Information

Indirect Identifiers		New Information
Sex	Year of Birth	Grade [%]
Male	1992	60
Male	1999	77
Male	1994	89
Female	2001	56
Female	1997	84
Male	2000	68
Male	2002	95
Female	1996	83
	1993	79
	2003	91
Male	1995	66

Which record belongs to Sally?

image:

This has been reproduced from the [De-identification Guidelines for Structured Data](#) published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

1 Manual annotation

CLINICAL NOTE

Encounter Date: 04/15/2024
Note Type: Progress Note

Subjective:
Patient reports mild headache and fatigue.
No fever or chills.

Patient Information

Patient Name: Alex Johnson
Date of Birth: 08/12/1978
MRN: MRN1234567
Address: 123 Maple St.
Springfield, IL 62704
Phone: (555) 123-4567

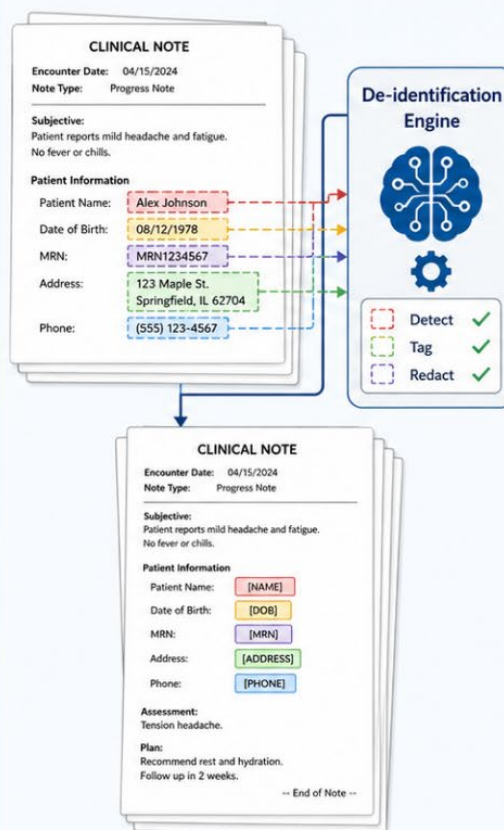
Assessment:
Tension headache.

Plan:
Recommend rest and hydration.
Follow up in 2 weeks.

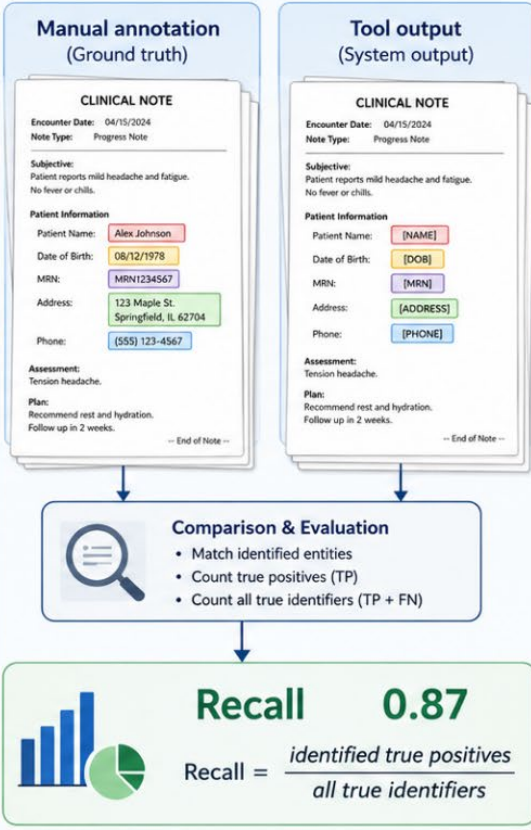
-- End of Note --



2 Tool-based processing



3 Compare outputs to compute recall



Evaluation framework assumptions

One Document = One Patient

Every document pertains to a single patient. If a document covers multiple patients, information may potentially be split into separate virtual documents.

Leak = Re-identification

Any undetected annotation (leak) is assumed to be a vulnerability for re-identification.

Caught = Protected

If an annotation is detected and redacted or re-synthesized, the probability of re-identification from that annotation is assumed to be zero.

Direct vs. Quasi-Identifiers

Direct identifiers (names, SSNs, MRNs) are treated separately from quasi-identifiers (dates, ZIP codes, cities) because their re-identification mechanisms differ.

Identifier detection metrics: Precision and recall



Contains identifiers in sentences

What should be protected

What the tool detected

How precision ($\frac{TP}{TP+FP}$) / recall ($\frac{TP}{TP+FN}$) is calculated

Example:

Gold: NAME = "Ana Liu"	EMAIL = "ana@example.com"	ID = "H-417"
Prediction: NAME = "Ana"	EMAIL = "ana@example.com"	

One output, multiple valid metric choices

Approach 1: Treat each identifier occurrence as independent

- Occurrence-level metrics evaluate individual gold standard and predicted identifier occurrences. A match can be defined using different rules:
 - Character-level
 - Partial-overlap-span level
 - Exact-span level
 - Full-containment-span level
 - Asymmetric containment-span level

Approach 1: Treat each identifier occurrence as independent

Character-level symmetric untyped precision and recall

Example 1: over-span + type mismatch

Clinical text: "Follow up with Dr. Ana Lee tomorrow."

Gold: PERSON = "Ana Lee"

Prediction: DOCTOR = "Dr.
Ana Lee"

Metric	TP?	FP?	FN?
Precision	7	4	
Recall	7		0

Example 2: under-span + type mismatch

Clinical text: "Spoke with Alicinia Chen."

Gold: PERSON = "Alicinia
Chen"

Prediction: NAME =
"Alicinia"

Metric	TP?	FP?	FN?
Precision	8	0	
Recall	8 of 21		5

Approach 1: Treat each identifier occurrence as independent

½ overlap-span level symmetric untyped precision and recall

Example 1: over-span + type mismatch

Clinical text: “Follow up with Dr. Ana Lee tomorrow.”

Gold: PERSON = “Ana Lee”

Prediction: DOCTOR = “Dr.
Ana Lee”

Metric	TP?	FP?	FN?
Precision	1	0	
Recall	1		0

Example 2: under-span + type mismatch

Clinical text: “Spoke with Alicinia Chen.”

Gold: PERSON = “Alicinia
Chen”

Prediction: NAME =
“Alicinia”

Metric	TP?	FP?	FN?
Precision	1	0	
Recall	13 of 21		0

Approach 1: Treat each identifier occurrence as independent

Exact-span level symmetric untyped precision and recall

Example 1: over-span + type mismatch

Clinical text: "Follow up with Dr. Ana Lee tomorrow."

Gold: PERSON = "Ana Lee"

Prediction: DOCTOR = "Dr.
Ana Lee"

Metric	TP?	FP?	FN?
Precision	0	1	
Recall	0		1

Example 2: under-span + type mismatch

Clinical text: "Spoke with Alicinia Chen."

Gold: PERSON = "Alicinia
Chen"

Prediction: NAME =
"Alicinia"

Metric	TP?	FP?	FN?
Precision	0	1	
Recall	0		1

Approach 1: Treat each identifier occurrence as independent

Full-containment-span level symmetric untyped precision and recall

Example 1: over-span + type mismatch

Clinical text: "Follow up with Dr. Ana Lee tomorrow."

Gold: PERSON = "Ana Lee"

Prediction: DOCTOR = "Dr.
Ana Lee"

Metric	TP?	FP?	FN?
Precision	1	0	
Recall	1		0

Example 2: under-span + type mismatch

Clinical text: "Spoke with Alicinia Chen."

Gold: PERSON = "Alicinia
Chen"

Prediction: NAME =
"Alicinia"

Metric	TP?	FP?	FN?
Precision	0	1	
Recall	0		1

Approach 1: Treat each identifier occurrence as independent

Containment-span level asymmetric untyped precision and recall

Example 1: over-span + type mismatch

Clinical text: "Follow up with Dr. Ana Lee tomorrow."

Gold: PERSON = "Ana Lee"

Prediction: DOCTOR = "Dr.
Ana Lee"

Metric	TP?	FP?	FN?
Precision	0	1	
Recall	1		0

Example 2: under-span + type mismatch

Clinical text: "Spoke with Alicinia Chen."

Gold: PERSON = "Alicinia
Chen"

Prediction: NAME =
"Alicinia"

Metric	TP?	FP?	FN?
Precision	1	0	
Recall	0 of 21		1

Approach 2: Treat each identifier unique value in a document as independent

Assumptions:

1. Within a document, every occurrence of the same identifier value is treated as carrying the same identifying information about the target.
2. If any occurrence remains exposed, the target may still be identified.
3. Additional exposed occurrences of the same value are not counted as additional information leakage.
4. Therefore, a unique identifier value is considered fully protected only when every occurrence in the document is detected.

All-or-nothing recall

$$\text{Recall_AON} = \frac{\# \text{ Fully detected gold unique values}}{\# \text{ Gold unique values}}$$

All-or-nothing precision

$$\text{Precision_AON} = \frac{\# \text{ Fully detected gold unique values}}{\# \text{ Predicted unique values}}$$

Why this matters for privacy risk:

- It directly measures whether an identifier remains leaked anywhere in the document, rather than how many individual occurrences the tool detects.
- For example, suppose a name “Ana Liu” appears three times in a document, but the tool detects only two occurrences. Occurrence-level recall is (2/3). However, one occurrence remains exposed, so the person can still be identified. In contrast, the all-or-nothing unique-value recall for “Ana Liu” is 0 because the unique value was not fully protected.

Micro- vs. Macro-average precision/recall

Micro-average

$$\text{Precision_micro} = \frac{\text{sum}(TP) \text{ across all documents}}{\text{sum}(TP) \text{ across all documents} + \text{sum}(FP) \text{ across all documents}}$$

$$\text{Recall_micro} = \frac{\text{sum}(TP) \text{ across all documents}}{\text{sum}(TP) \text{ across all documents} + \text{sum}(FN) \text{ across all documents}}$$

Interpretation: every evaluated unit (either an identifier occurrence or an identifier unique value) has equal weight.

Macro-average

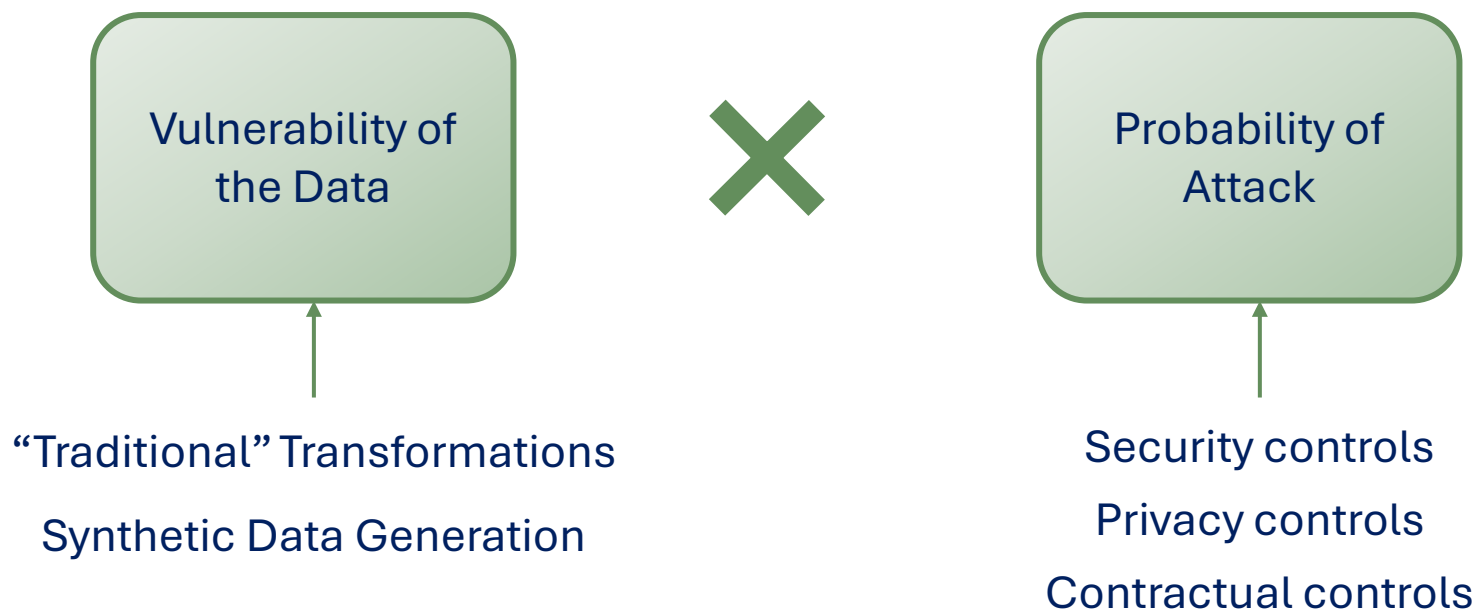
$$\text{Precision_macro} = \text{average}(\text{Precision_doc_1}, \dots, \text{Precision_doc_N})$$

$$\text{Recall_macro} = \text{average}(\text{Recall_doc_1}, \dots, \text{Recall_doc_N})$$

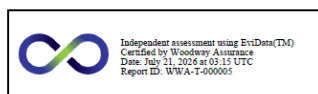
Interpretation: each document contributes the same weight, regardless of the number of identifiers in a document.

Macro-average is often less common because some documents could have zero or very few identifiers, making document-level scores unstable or undefined.

Multiple Levers to Manage Risk



Risk Assessment Report



EviData™
powered by
Woodway Assurance

Text Re-identification Risk Report

Data Custodian: local-demo
Data Recipient: local-demo
Date: 21 July 2026
Report ID: WWA-T-000005
Document set: corpus.zip
Documents assessed: 2
Generated by: EviData™ by Woodway Assurance Ltd
(version 1.0.0, contact: info@woodway-assurance.com, www.woodway-assurance.com)



The Identity Disclosure Risk is **HIGH**.

Disclaimer: This assessment is based on the documents provided by the Data Custodian and the assumptions stated herein. A **LOW** verdict indicates that the assessed residual identity disclosure risk lies below the thresholds accepted by the Data Custodian; it is not a warranty of anonymity. A re-assessment is recommended every two years, and is required sooner if the source content, the reference model, the evaluated de-identification tool, or the thresholds change in a material way.

- A report is produced automatically based on a sample of documents that are considered during the evaluation stage
- The report covers the risk assessment methods and results, as well as the controls that need to be put in place by the data custodian and the data recipient

