

De-identifying company workspace exports in the data holder's browser: design and benchmark results

Yuma Heymans

AIData.Broker, <https://aidata.broker> | hello@aidata.broker | 18 October 2026

Abstract

Company records such as chat workspaces, mailboxes, documents and tables are proposed as training data for language models, and they carry identifiers of employees, customers and other organisations. This report describes a rule-based de-identification engine that runs entirely in the data holder's browser, so that raw records, the people learned from them and the key behind every pseudonym stay on the holder's computer. It combines the people an export names in its own structure, pattern detectors validated by check digits, cues and a public-domain name lexicon, and the holder's own lists. It writes stand-ins from a keyed HMAC-SHA-256, scoped to each file by default, holds every table row to a group of at least 11, screens content for special categories, and rereads its output with a second check that has no memory of the first pass. On the held-out split of the Text Anonymization Benchmark (127 court judgments) it masks every mention of 93.4% of direct identifiers. On the six human-annotated exports of PrivacyBench v2.5, not inspected while the rules were written, it reaches 86.8% F1 (precision 88.1%, recall 85.6%) using export structure, and 81.9% F1 on text alone. It is weakest on organisation names (49.9% recall), links (61.9%), account numbers (68.3%) and street addresses (87.4%). We report the protocol, the published comparison table and the threats to validity.

Keywords: de-identification; pseudonymisation; personally identifiable information; named entity recognition; workplace data; k-anonymity; privacy engineering; training data; benchmark

1. Introduction

A company's working records are text written by and about people. A month of a chat workspace carries the names of colleagues, their handles and user ids, their phone numbers in signatures, customer account numbers pasted into threads, links to private documents, and the names of clients and suppliers. When such records are proposed as training data for language models, the first technical question is how much of that can be removed before anyone outside the company reads the text, and how well the removal works.

Where the removal happens matters as much as how. If a recipient scrubs records after receiving them, the recipient has already received what was meant to be removed. NIST's guidance on de-identifying government data states the requirement plainly: "The software does not leak identifying information in expected or unexpected ways" [1, section 5.3]. The European Data Protection Board's guidelines on pseudonymisation add that only "possession of the secret information should enable the computation of the pseudonym given the identifier" [2, para 88]. Practice varies. Scale AI's data partnership page states that it "anonymizes everything inside your environment, before any data is extracted" [3]. In a 2026 bankruptcy sale of an airline's records, the court-appointed consumer privacy ombudsman described a structure in which "a key component is having an independent expert conduct the deidentification" [4].

This report describes an engine built for that placement: it runs in a Web Worker in the browser of the person who holds the records, and what can leave the computer is the rewritten text and a report of counts. It makes three contributions.

1. A design for de-identifying workspace exports (chat, email, documents, tables) that runs without a server and without a learned model, built from layered rule-based evidence, keyed pseudonyms scoped per file, a group-size test for tables, a content screen and an independent second check.
2. An evaluation on two public benchmarks, each with a part that was held out while the rules were written: the Text Anonymization Benchmark (TAB) of court judgments [5, 6] and PrivacyBench v2.5 of synthetic company exports [7, 8].
3. An account of where the engine fails, by kind of identifier and by export, with the numbers as measured.

Section 2 gives background, Section 3 the design, Section 4 the protocol, Section 5 the results, Sections 6 and 7 the weaknesses and threats to validity, and Section 8 reproducibility. Regulations are cited only to explain design choices; nothing here is legal advice.

2. Background

Terminology. NIST advises "avoiding the phrase anonymization", because "there is no technical mechanism to ensure that de-identified data cannot be re-identified" [1, p. 51], and the EDPB's draft guidelines warn against calling data "anonymous" or "de-identified" "if individuals are actually still identifiable" [9, para 40]. This report follows NIST: the engine de-identifies, and its output is described by what was measured.

Rules and learned models. NIST found that "rule-based systems excelled at finding certain kinds of information (e.g., SSNs and phone numbers), while the trainable systems worked better for other kinds of data" [10, p. 31]. Names are the hard case for rules, and "direct identifiers cannot be limited to predefined categories of entities" [6].

Measuring protection, not mentions. TAB introduced an entity-level measure: "Whenever an entity is mentioned several times in a given document, it only makes sense to view this entity as 'protected' if all of its mentions are masked" [6, section 6.1]. On TAB's test set, Presidio with default settings protected 0.460 of direct identifiers by this measure, and a zero-shot neural tagger 0.940 at a precision of 0.441 [6, Table 7]. A mention-level recall can look acceptable while whole people stay exposed.

Workplace exports. PrivacyBench, published by Tonic.ai, holds one month of synthetic workspace exports (Slack, email, Drive documents) for 21 fictional protagonists at different companies, with ten labelled kinds of identifier and human annotation of six exports [7]. Its card states that Tonic "built this benchmark and the system that scores highest on it" and designed it so that anyone can compare a system with the published scores; the scorer is open source [8].

Pseudonyms and linkage. An unkeyed hash is reversed by hashing every candidate value, as happened to New York taxi data [10, p. 17], and HHS treats a code from an unkeyed hash as identifying [11]. A pseudonym consistent across a corpus lets an attacker collect what one person said in many places: "an attacker who reconstructs a pseudonym in many sentences is more likely to re-identify the person" [12], and the AOL search logs were re-identified through such a link [10, p. 12].

Tables. Sweeney estimated that 87% of the US population was likely unique on five-digit ZIP code, gender and date of birth [13]. EMA and Health Canada set an acceptable re-identification probability of 0.09 for public clinical reports [14], a smallest group of 11, and EMA advises using the maximum risk for

public release [15]. HIPAA's safe harbour truncates ZIP codes to three digits, writes 000 for areas of 20,000 people or fewer, and groups ages over 89 [16].

3. System design

3.1 Placement and threat model

The engine is a TypeScript module with no third-party dependencies; file readers sit outside it. In the deployed tool [17] it runs in a Web Worker on the holder's computer. The holder selects exports; the worker reads each file through local readers, rewrites it, and returns the rewritten files, a report and a review list to the page. The raw records, the roster of people, the review list and the pseudonym key stay in the browser. The report holds counts only: records and words read, identifiers replaced by kind, people given a stand-in, table group sizes, special-category signals, review results and the second check's findings. It contains no text, names or values, so it can travel with the output as evidence.

The adversary considered is a recipient of the rewritten text who wants to identify a person in it, using the text itself, public information and other data. The engine addresses the identifiers that text carries on its surface. It does not claim to defeat identification by writing style or by combinations of ordinary facts; those residual risks are addressed by human review and independent testing outside the engine, as discussed in Section 7.

3.2 Two passes over an export

Company exports name their people in structured places before they mention them in prose: a Slack export's users.json, an email's From, To and Cc headers, a ticket's reporter and assignee, a commit's author, a CRM's contact columns, a transcript's speaker labels. The engine reads those first. In pass one, the format readers add every person they find to a roster with their full name, first and last names, email addresses, handles and display names. In pass two, every record is rewritten, and every mention of a roster person (full name, first name, surname, email address, Slack mention or bare username) takes that person's single stand-in. This is why "Jane", "jane.doe@acme.com", "<@U024BE7LH>" and "Jane Doe" all become the same [PERSON_3].

For each piece of text the engine runs, in order, the pattern detectors, the name layer, the organisation layer and the holder's lists; it then resolves overlaps (the longer find wins, and at equal length a fixed priority by kind), replaces each find, and updates the counts, the local review list and the content screen.

3.3 The identifier catalogue

The engine's catalogue has 17 kinds of identifier, each with a treatment. Nine are replaced by a pseudonym that keeps who-is-who, seven by a typed tag with no link, and one (secrets) is removed and marked.

Kind	Treatment	How it is found (summary)
Name of a person	pseudonym	roster, cues, name lexicon
Email address	pseudonym	syntax; shared mailboxes and example domains kept
Phone number	pseudonym	grouped or international runs of 9 to 15 digits
Postal address or postcode	tag	house number and street; postcodes by shape
Date of birth	tag	a date next to a birth word
National identity or tax number	tag	format and check digit (Table 2), or after its label
Card number	tag	13 to 19 digits passing Luhn, with a card prefix
Bank account number	tag	IBAN with mod-97 check; numbers after account words
Customer, employee or account number	pseudonym	label, kind prefix, learned shape, contact line
IP address	pseudonym	public IPv4 and IPv6
Device id	pseudonym	MAC shape; IMEI and advertising ids after their name
User id or handle	pseudonym	chat mentions, @handles, labelled usernames
Personal or internal link	tag	profile hosts, private file and workspace links
Precise location	tag	coordinates with four or more decimals
Name of an organisation	pseudonym	export domains, legal and trade words, listed companies
Password, key or token	removed	published key formats, private keys, JWTs
Term the holder listed	pseudonym	whole-word match

Table 1. The engine's identifier catalogue (engine version 2026-10-16.4).

The detectors err toward flagging. A false alarm costs the holder a second look at one word; a miss sends a person's details.

3.4 Pattern detectors and check digits

A detector is a pattern for the shape, then the identifier's own validity rule wherever one exists, and a keyword nearby where the shape alone would match ordinary numbers. Patterns and check digits follow Microsoft Presidio's recognisers (MIT licence) [18] and each issuing authority's published algorithm; secret formats follow gitleaks' rules (MIT licence) [19].

Identifier	Validity rule implemented
Payment card	Luhn checksum and card prefix
IBAN	ISO 13616 rearrangement and $\text{mod } 97 = 1$
Dutch BSN and RSIN	the 11-proof, $(9a + 8b + \dots + 2h - i) \text{ mod } 11 = 0$
German tax id	ISO 7064 MOD 11,10
French NIR	key = $97 - (n \text{ mod } 97)$, with Corsican 2A and 2B mapped
Belgian national number	$97 - (n \text{ mod } 97)$, with the post-2000 prefix
Norwegian national number	two weighted mod-11 check digits and a valid date
Finnish personal identity code	check character from a 31-symbol alphabet
Spanish DNI and NIE	check letter from a table indexed by $n \text{ mod } 23$
Italian codice fiscale	check letter from odd and even position tables
Polish PESEL	weights 1, 3, 7, 9 and $(10 - \text{sum mod } 10) \text{ mod } 10$
Indian Aadhaar	Verhoeff checksum
Brazilian CPF	two mod-11 check digits
Singapore NRIC and FIN	weights 2, 7, 6, 5, 4, 3, 2 and a letter table per prefix
Australian TFN	weights 1, 4, 3, 7, 5, 8, 6, 9, 10, $\text{sum mod } 11 = 0$
UK NHS number	mod 11 over weights 10 to 2
US employer identification number	two-digit prefix from the IRS list of valid campus prefixes [20]
Canadian SIN	Luhn checksum, excluding leading 0, 8 and 9
Swedish personal number	Luhn checksum and a valid date
IMEI	Luhn checksum
Bitcoin address	Base58Check double SHA-256, or Bech32 and Bech32m checksums

Table 2. Validity rules in the pattern layer. US SSN and ITIN are matched by format and the issuer's excluded ranges, UK National Insurance numbers by format and excluded prefixes, Danish CPR numbers by a valid date, and Indian PAN by format; passports, driving licences and other national numbers are taken after their label.

Keyword anchoring is used where a valid shape is common: a Dutch BSN is any nine digits that pass the 11-proof, so the detector requires "BSN" nearby. The phone test rejects bare digit runs without separators, dates, version numbers, amounts, decimals and a year followed by a number (a case reference).

Two mechanisms extend the pattern layer within a run. An identifier found with its label teaches its shape: once "PFE-208814" is labelled as an employee id, later codes with the same prefix and number of digits in that export are taken too. A number of six or more digits found with a label (an order, an account, a national id) is found again wherever it recurs unlabelled in the same export. A code on the same line as a person's email address or phone number is taken as that person's id.

3.5 Names of people

Names have no shape a pattern can check, so the name layer combines four kinds of evidence, strongest first.

1. **The roster.** People the export lists are found wherever mentioned, in any case.
2. **Cues.** A greeting or sign-off ("Hi Xiulan", "Thanks, Marco"), a title ("Dr Okafor"), a byline ("assigned to", "reported by"), a speaker label at the start of a line, a form's field label ("Name:", "Signature:"), or chat context ("ask rob", "jordan said") mark a name even when no list holds it.
3. **The lexicon.** A capitalised given name followed by a surname. The lexicon (built 15 October 2026) draws on US Social Security baby names, US Census 2010 surnames and Wikidata names, with Project Gutenberg word frequencies marking common words [21, 22, 23, 24]: 20,745 given names, 20,148 surnames, 21,436 common words and 621 ambiguous given names.
4. **The holder's lists.** Terms to remove (clients, codenames) and terms to keep (the holder's own company and products).

A given name that is also a common word, a month or a place ("Will", "May", "Paris") is taken only with other evidence, and determiners, place words and organisation endings ("Harvey AI") stop a capitalised word from being read as a person. Mentions by surname alone, in non-Latin scripts or by nickname are found only through the roster or the holder's list. Capitalised words no rule explains are not replaced but listed for the holder to check.

3.6 Names of organisations

Organisation names are not personal data on their own, but they are often how a person is found again ("the treasury contact at" a named firm is one person), and a client's name beside its terms is often another company's confidential information. The engine therefore replaces them by default, and the holder keeps the ones a recipient may see. Evidence comes from:

1. the export's own email domains: hudsonandvine.com makes "Hudson & Vine" an organisation wherever it is written (free mailbox domains teach nothing);
2. a legal word (Inc, LLC, GmbH) or trade word (Logistics, Studios) after a capitalised name, after which the name alone is learned for the run;
3. the 8,403 companies in the SEC's company_tickers.json (built 15 October 2026) [25], where a name that is also a common word or a person's name is taken only when capitalised inside a sentence;
4. the holder's lists.

Where the name layer and the organisation layer overlap on a name the roster does not hold ("Raymond James", "Morgan Stanley"), the organisation takes the whole span. A name the roster holds stays a person. Departments, products and brands are not treated as organisations. Links to a domain the run has learned as an organisation's are replaced as links.

3.7 Keyed stand-ins scoped per file

Every pseudonym number comes from a keyed hash. At the start of a run the browser generates a 256-bit secret with the Web Crypto API. For each value, the engine computes HMAC-SHA-256 over its kind and canonical form under that secret, and uses the first 12 bytes as the key of a table that maps keyed hashes to sequential numbers; the table holds hashes and numbers, not names. A new value takes the next number for its kind: [PERSON_1], [PERSON_2] and so on.

The holder may save the key (the secret and the table) to get the same stand-ins in a later run; without it, the link ends when the browser tab closes. A new key for each recipient means two recipients' copies cannot be joined through their stand-ins.

By default, the identity hashed for a stand-in is the value prefixed with the identifier of the file being rewritten (a Slack channel-day, a mailbox, a table). A person therefore keeps one stand-in within a file and takes a different one in the next, so the clues about them cannot be gathered across the archive through one pseudonym. An archive-wide scope is available for uses that need who-is-who across conversations; a delivery made with it carries an open check stating that linkage across files must be tested independently.

A person's email address and handle share the person's number (person_3@redacted.invalid). An optional surrogate mode writes realistic stand-ins instead: a fictional name chosen by a second keyed hash, an address at example.com, a phone number in the 555-01 range reserved for fiction. Surrogates let a missed name hide among stand-ins that look the same [1, p. 51]. Tags are the default for samples checked again on arrival, because a tag is recognisable to the second check.

3.8 Tables

Tables (CSV, spreadsheets, JSON rows) are handled column by column. Each column's role is classified once from its header and a sample of values: a direct identifier, an organisation, a quasi-identifier (postcode, birth date, age, gender, job title, city), free text, or other. Direct identifiers take the stand-in the person has elsewhere in the export, free text is rewritten like prose, and other columns are still read for identifiers.

With generalisation on (the default), quasi-identifiers are coarsened. US ZIP codes are cut to three digits, and 23 three-digit prefixes covering 20,000 people or fewer are written as 000: the 17 that HHS lists from the 2000 Census [11] and six added by a recomputation from the 2010 Census. UK postcodes are cut to their outward code, Canadian ones to the forward sortation area, Dutch ones to the first two digits. Birth dates become the year. Ages become five-year bands, with 90 and over as one group.

The combination of quasi-identifier values in each row is then counted. A delivery may contain no row whose combination is shared by fewer than 11 rows in its table, the EMA and Health Canada threshold of 0.09 [14] applied to every row rather than on average, as EMA advises for public release [15]. In a sample sent only for measurement, groups smaller than 5 are flagged; the UK ICO cites $k = 5$ as the NHS standard for strong anonymisation [26].

3.9 Content screen and risk signals

Each record is also screened by keyword patterns for nine special categories (health, children, sex life or orientation, religious or political beliefs, trade union membership, racial or ethnic origin, criminal records, biometric or genetic data, a person's finances) and three kinds of content that removing names does not make shareable (HR matters, legal advice, and third parties' confidential material). Matches are counted per record and flagged for the holder to read; the engine does not remove or pass such records on its own judgment. A run with fewer than five distinct people is flagged, since writing style and role can identify members of a very small group whatever is replaced.

3.10 The second check and the verdict

Every rewritten record is read again by a function with the same detectors, lexicon and organisation layer, but no roster and no memory of the first pass; the holder's terms to keep (and, in surrogate mode, the run's surrogates) are excluded. In the deployed tool a copy cannot be downloaded while this check finds anything; its finds are shown in context so the holder can add them to the terms to remove and run again.

The report is then read against a policy for a sample (sent for measurement) or a delivery. Each check passes, stays open (a person still has to act) or fails:

- no direct identifier left (the second check found none);
- a person read records picked at random and found none: with nothing found in n records, the share still holding an identifier is below $1 - 0.05^{(1/n)}$ at 95% confidence, so a delivery requires 300 records read, giving a bound under 1.0%, and a copy smaller than the required reading is read whole;
- every table row in a group of at least 11 (delivery) or flagged below 5 (sample);
- special categories and content screened (open while any record is flagged);
- linkage across files tested (open for a delivery made with the archive-wide scope);
- a group of at least five people (open when fewer appear);
- certification by an independent provider before a delivery (open by design, since the engine cannot certify its own output).

When a sample reaches the server, it is read again by the same rules with no roster and by a language model that reads 150 records chosen at random; what either finds is replaced before the sample is measured, and the text is discarded when the request ends [17].

4. Evaluation protocol

4.1 Benchmarks

TAB [5, 6] consists of 1,268 English-language judgments of the European Court of Human Rights, in which legal annotators marked every identifier, its type, and whether it must be masked (as a direct or quasi-identifier). It is distributed under the MIT licence. We use its development split (127 judgments) as the held-out set and also report its test split (127 judgments), which was read while the rules were written and is therefore not held out.

PrivacyBench v2.5 [7] is distributed under CC BY 4.0. We used the dataset revision a79d9e80c5ee68f5dd79da8452e5eb1e1ba97a1e of 15 October 2026. Six of its 21 exports carry human annotation: daniel_lincoln_financial_group, isabella_chipotle, megan_donovan_eli_lilly, morgan_morgan_stanley, nora_caterpillar and renee_tyson_foods, with 5,198 messages, 656 document pages and 25,855 gold spans. These six are our held-out set. The other 15 exports, which carry only the generated answer key, are our development set: misses on them were read and rules improved. The card notes that the generated key lists the seeded cast of characters and their organisations and is less complete than the human annotation [7], so development recall is measured against an incomplete key and precision is not computed on it.

4.2 Held-out discipline

By the protocol recorded in the two measurement scripts, the text and misses of the held-out sets (the TAB development split and the six human-annotated PrivacyBench exports) were not read while rules were built; misses were read only on the TAB test split and the 15 PrivacyBench development exports. Aggregate held-out scores were, however, computed at more than one engine version during development, so selection among rule changes on aggregate held-out scores cannot be excluded.

4.3 Configurations

All runs use the engine's default options: tag stand-ins, file scope, organisations replaced, generalisation on, no holder lists. Engine version 2026-10-16.4 and lexicon version 2026-10-15 were measured on 16 October 2026.

On TAB, the engine runs with no roster, since court judgments have no structured list of people; only its dictionaries, cues and patterns are measured.

On PrivacyBench, one engine is built per export, as the tool builds one per run. In the default configuration (**with roster**), the export's own structure files (Slack users.json and day files, every email's headers, the Drive listing) pass through the tool's format readers to build the roster, and each message body and each document's extracted text is then read by the engine. The answer key's character list is not read. In the ablation (**text only**), no roster is built. In both, documents are read as the benchmark's extracted text, as were all engines in the published comparison except the top row [7].

4.4 Metrics

TAB. Following the benchmark [6, section 6.1], a direct identifier is protected only if every one of its mentions is masked; entities are counted per annotator, over all annotators. We report entity-level and mention-level recall for direct identifiers, entity-level recall for other people (marked as quasi-identifiers) and for all quasi-identifiers, and a precision proxy: the share of masked characters inside spans the first annotator marked for masking. The benchmark's own precision is token-level and weighted, so the proxy is close to it but not equal. A title before a name is not counted as part of the name, because the engine keeps titles.

PrivacyBench. We use the benchmark's matching rule (score_ner.py): a gold span is found when a prediction with the same label overlaps it, and a prediction is correct when it overlaps a gold span of its label; ID_NUMBER predictions count for EMPLOYEE_ID and ACCOUNT_NUMBER, as in the scorer's alias table. A person's name is split as the benchmark's baselines split theirs (first word given, last word surname). Messages are annotated with five labels (given name, surname, email, username, organisation) and documents with all ten, so predictions in messages are restricted to those five. We report pooled precision, recall and F1; exact-boundary recall; per-label and per-export results; and the counts without organisations ("identifiers of people"). The benchmark's published scorer [8], run on the same predictions, agreed with our implementation on pooled precision and recall to within 1e-9.

5. Results

5.1 Court judgments (TAB)

Measure	Dev split (held out)	Test split (read during development)
Direct identifiers, every mention masked	93.4% (1,080 of 1,156)	95.9% (1,110 of 1,157)
Direct identifier mentions masked	92.9% (1,347 of 1,450)	97.3% (1,568 of 1,612)
People marked as direct identifiers	93.3% (533 of 571)	94.8% (545 of 575)
Other people (marked as quasi-identifiers)	88.4% (2,140 of 2,420)	85.3% (2,079 of 2,436)
All quasi-identifier entities	24.4% (3,979 of 16,299)	18.9% (3,071 of 16,291)
Precision proxy (masked characters inside must-mask spans)	88.3%	89.5%
Share of all characters masked	2.7%	2.4%

Table 3. The engine on TAB, no roster, engine 2026-10-16.4.

For reference, the benchmark's authors measured two systems on the test split [6, Table 7]:

System (as measured by Pilán et al. on the test split)	Direct identifiers, entity level	Quasi-identifiers, entity level	Precision
Microsoft Presidio (default)	0.460	0.758	0.761
Neural tagger (RoBERTa on OntoNotes v5, zero-shot)	0.940	0.874	0.441

Table 4. Reference systems on TAB as published. Their precision is the benchmark's token-level weighted measure, not our character-level proxy, and their split differs from our held-out split.

On the held-out split the engine protects 93.4% of direct identifiers while masking 2.7% of the text, 88.3% of it inside must-mask spans. Its low quasi-identifier recall (24.4%) is a design choice: dates and places in prose are left by default and handled by the table checks, the review and the holder's lists.

5.2 Workplace exports (PrivacyBench): the comparison table

Engine	Precision	Recall	F1
Textual Graph Synthesis	90.9%	94.3%	92.5%
Textual via SDK	91.3%	92.9%	92.1%
Claude Opus 5	92.1%	84.9%	88.3%
Claude Sonnet 5	92.1%	82.4%	87.0%
GPT-6 astra	88.6%	85.0%	86.8%
This engine, with roster	88.1%	85.6%	86.8%
GPT-5.6 sol	90.2%	82.3%	86.1%
GPT-5.6 terra	92.4%	78.4%	84.8%
Presidio	79.0%	87.1%	82.8%
This engine, text only	89.7%	75.4%	81.9%
GLiNER2	83.7%	68.0%	75.1%

Table 5. Held-out human-annotated exports, overlap matching. Rows other than this engine's are copied from the PrivacyBench v2.5 card's NER table as read at revision a79d9e8 (15 October 2026), with engine

names as the card gives them; we did not rerun them.

With its default use of export structure, the engine's F1 (86.8%) is level with the GPT-6 astra row at the precision reported, below the two Textual rows and the two Claude rows, and above the remaining rows. On text alone, its F1 (81.9%) is below Presidio's and above GLiNER2's.

One qualification matters for reading Table 5. The card states that every row except Textual Graph Synthesis "was run on pre-extracted text (the message bodies and the extracted text of the documents)", and that Graph Synthesis also uses the structure of the export, such as email headers and a Slack export's users file [7]. The engine's default configuration uses that structure too. The fair comparison with the pre-extracted-text rows is therefore the engine's text-only row; the roster row compares with Graph Synthesis, which additionally extracts document text itself.

5.3 By kind of identifier

Label	Gold spans (held out)	Found, with roster	Found, text only	Development recall (15 exports, generated key)
Given names	10,034	94.7% (9,504)	70.2% (7,048)	96.5% (33,670 of 34,904)
Surnames	5,204	96.6% (5,029)	82.6% (4,301)	98.2% (20,403 of 20,771)
Email addresses	2,453	99.7% (2,445)	99.7% (2,445)	99.6% (10,860 of 10,908)
Usernames and Slack ids	451	99.3% (448)	90.9% (410)	98.5% (1,229 of 1,248)
Phone numbers	1,036	100.0% (1,036)	100.0% (1,036)	99.9% (4,696 of 4,699)
Street addresses	419	87.4% (366)	87.4% (366)	98.9% (1,215 of 1,229)
Employee ids	295	97.3% (287)	97.3% (287)	94.1% (952 of 1,012)
Account numbers	167	68.3% (114)	68.3% (114)	63.6% (292 of 459)
Links	21	61.9% (13)	66.7% (14)	68.6% (236 of 344)
Organisations	5,775	49.9% (2,879)	60.0% (3,465)	64.0% (14,229 of 22,220)
All labels	25,855	85.6% (22,121)	75.4% (19,486)	89.8%

Table 6. Recall by label. Held-out figures are against the human annotation; development figures are against the generated key, which is less complete.

Without organisations, over the 20,080 gold spans of people's identifiers, the engine with its roster reaches precision 90.6%, recall 95.8% and F1 93.1% (19,242 found, 21,587 predicted). On text alone the same subset gives precision 93.2%, recall 79.8% and F1 86.0%.

5.4 By export

Export (held out)	Gold spans	Precision	Recall	F1	Exact-boundary recall	F1, text only
daniel_lincoln_financial_group	4,488	89.8%	85.1%	87.4%	79.3%	82.5%
isabella_chipotle	4,098	82.7%	92.9%	87.5%	85.6%	81.4%
megan_donovan_eli_lilly	4,142	88.2%	90.4%	89.3%	84.9%	84.2%
morgan_morgan_stanley	4,427	88.0%	84.2%	86.1%	74.8%	80.0%
nora_caterpillar	4,177	92.7%	76.3%	83.7%	70.3%	79.9%
renee_tyson_foods	4,523	88.3%	84.8%	86.5%	78.6%	83.4%
Pooled	25,855	88.1%	85.6%	86.8%	78.8%	81.9%

Table 7. Per-export results with roster, and F1 without.

F1 ranges from 83.7% to 89.3% across the six exports. Precision is lowest on isabella_chipotle (82.7%), where recall is highest (92.9%); recall is lowest on nora_caterpillar (76.3%), where precision is highest (92.7%).

5.5 What the roster contributes

The roster raises pooled recall from 75.4% to 85.6% and F1 from 81.9% to 86.8%, at a cost of 1.6 points of precision. The gain is where expected: given names rise from 70.2% to 94.7% and surnames from 82.6% to 96.6%, because colleagues are often mentioned by first name alone ("thanks jane"), which no dictionary rule can safely take; usernames rise from 90.9% to 99.3%. Labels found by patterns do not change.

Organisations move the other way, from 60.0% to 49.9%. The design gives a roster name precedence over an overlapping organisation find (Section 3.6), so a plausible mechanism is that roster names coincide with words inside organisation names. We have not verified this span by span.

5.6 Boundaries

Overlap matching credits a prediction that covers part of a gold span. Exact-boundary recall is 78.8% pooled with the roster (89.5% for people's identifiers) and 68.7% on text alone; the gap to overlap recall (85.6%) is the share of gold spans touched with different boundaries. A partial cover can leave part of an identifier in clear text, one reason the second check reads the output again.

6. Weaknesses as measured

The engine's failures are concentrated in four labels. We state each with its numbers and the cause as far as we understand it.

Organisation names: 49.9% recall (2,879 of 5,775). This is the weakest label by a wide margin and the largest single source of missed spans: 2,896 of the 3,734 missed gold spans are organisations. Recall ranges from 42.8% (megan_donovan_eli_lilly) to 61.8% (isabella_chipotle) across exports, and is 64.0% on the development exports against the generated key. Organisations are often written as brands or short forms without a legal or trade word, as department names, or as names not listed with the SEC. The engine deliberately does not treat departments, products and brands as organisations, while the benchmark's

annotation may, so part of the gap is a difference in definition and part is plain misses. The people-only figures in Section 5.3 isolate the identifiers of people.

Account numbers: 68.3% recall (114 of 167). The pattern layer finds account numbers through a label, a recognisable kind prefix, a learned shape or a contact line. Account numbers written with no word beside them, in a shape not seen labelled earlier in the export, are missed. The weakness is uneven: 32.6% (14 of 43) in `daniel_lincoln_financial_group` against 98.1% (53 of 54) in `renee_tyson_foods`. Development recall is similar (63.6%, 292 of 459), so this is not an artefact of the held-out sets.

Links: 61.9% recall (13 of 21). The held-out sample is small, so the estimate is wide; development recall is 68.6% (236 of 344). The engine replaces profile links, private file and workspace links, links carrying an email or token, and links to learned organisation domains; internal links to other hosts fall outside those categories.

Street addresses: 87.4% recall (366 of 419). Recall ranges from 73.9% (82 of 111) in `isabella_chipotle` to 100.0% in `nora_caterpillar`. The address patterns require a house number and a street type, or a postcode shape, and miss addresses written in other forms. Development recall is 98.9%, so the held-out gap suggests address forms in the human-annotated exports that the development exports did not show.

Names without a roster. On text alone, 29.8% of given names are missed. Workspace prose mentions colleagues by lower-case first names and nicknames that the lexicon rules do not take without corroboration. Exports without structure files, such as a pasted thread or a single document, should be expected to behave like the text-only row.

Precision. Pooled precision is 88.1%, and 82.7% in the lowest export. Over-masking costs utility rather than privacy, but a recipient loses text it could have used.

7. Threats to validity

Synthetic data and a shared generator. PrivacyBench's exports are fully synthetic [7]. Held-out and development exports come from the same generation process, so rules written from development misses may transfer to the held-out exports better than to real company exports, whose formats, languages and habits vary more.

Aggregate scores seen during development. As stated in Section 4.2, aggregate held-out scores were computed at more than one engine version. This is a weaker form of leakage than reading held-out text, but it is not none.

The comparison table. Other rows are the benchmark authors' published numbers; the authors build the top-scoring system and say so [7]. We did not rerun them. Our row is self-reported, though computed with the authors' scorer, and our default configuration uses export structure that most rows did not receive (Section 5.2).

Cross-split comparison on TAB. Pilán et al.'s reference figures are on the test split [6]; our held-out figures are on the development split. Our test-split figures are not held out. Our precision proxy is character-level and differs from the benchmark's token-level weighted precision.

Matching rules. Overlap matching is lenient. Exact-boundary recall is lower (78.8%). In messages, predictions were restricted to the five labels the benchmark annotates there, which removes possible false positives of other kinds from the count.

Language and format coverage. Both benchmarks are in English. The engine has rules for Dutch, German, French, Spanish and Italian address forms and several national identifiers that these benchmarks do not exercise, and those rules are unmeasured here. Documents were read as the benchmark's extracted text, so the tool's own document readers (and any text extraction errors they make) are not measured.

Detection is not re-identification risk. Both benchmarks measure whether marked spans were masked. They do not measure whether a person can still be identified. Long texts by one author can be attributed by style; Narayanan et al. report correct identification "in over 20% of cases" among 100,000 authors [27]. A model trained on scrubbed text can repeat what the scrub missed, and "scrubbing is imperfect" [12]. Rocher, Hendrickx and de Montjoye estimate that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes [28]. These are the reasons the design adds the content screen, the group-size test, a human review with a stated bound, and independent testing before any delivery; none of them is measured by the benchmarks.

Conflict of interest. The author leads the organisation that built and operates the engine.

8. Reproducibility

Versions. Engine 2026-10-16.4; name lexicon 2026-10-15 (20,745 given names, 20,148 surnames, 21,436 common words, 621 ambiguous given names); listed companies 2026-10-15 (8,403 names); measured 16 October 2026. PrivacyBench v2.5 at revision a79d9e80c5ee68f5dd79da8452e5eb1e1ba97a1e (15 October 2026). TAB from its public repository [5].

Scripts. Two scripts in the engine's source tree perform the measurements and write the result files that the public standard page reads [17]:

- apps/app/scripts/deid-benchmark.mjs downloads the TAB development and test splits, runs the engine with no roster on each judgment, and writes benchmark.json with the counts in Table 3.
- apps/app/scripts/deid-privacybench.mjs clones the PrivacyBench repository without its large files, builds one engine per export, runs the held-out exports with and without the roster and the development exports with it, writes the predictions in the benchmark's format, and writes benchmark-workplace.json with every count in Tables 5 to 7. With the flag --official, it also clones the benchmark's scorer [8], scores the same predictions with it, and records whether the two agree.

Inputs. Both benchmarks and the PrivacyBench scorer are public [5, 7, 8], and the card documents the command that scores any engine's predictions against the human annotation [7]. The engine runs in a current browser in the deployed tool [17].

Results. The figures in this report are copied from the two result files as written on 16 October 2026, rounded to one decimal place. The only figures derived by hand are differences between them: missed spans in Section 6 (gold minus found), the share of given names missed on text alone, and the precision difference in Section 5.5.

9. Conclusion

A rule-based engine that runs in the data holder's browser reaches, on held-out workplace exports, an F1 level with one frontier-model pipeline in the benchmark authors' published table and above two others, when it uses the structure the exports already carry. The roster of people learned from that structure is the most important single component: it raises recall on given names from 70.2% to 94.7%. Without it, the engine's F1 (81.9%) is slightly below Presidio's (82.8%).

The engine is not finished. It finds 95.8% of the identifiers of people in the held-out exports, but 49.9% of organisation names, 68.3% of account numbers, 61.9% of links and 87.4% of street addresses. Those failures are measured, located by export and by kind, and published with the engine version that produced them. Detection figures are also not a measure of re-identification risk, which is why the design places a second check, a human review with a stated statistical bound, a group-size test for tables, a content screen and an independent test between the engine's output and any recipient.

The work that follows from these results is specific: organisation names written as brands and departments, unlabelled account numbers, internal links, address forms, and an evaluation on non-English exports.

Competing interests

The author is the chief executive of AIData.Broker, which built and operates the engine described in this report.

References

- [1] S. L. Garfinkel et al., "De-Identifying Government Data Sets," NIST Special Publication 800-188, National Institute of Standards and Technology, September 2023. <https://doi.org/10.6028/NIST.SP.800-188> (accessed 18 October 2026).
- [2] European Data Protection Board, "Guidelines 01/2025 on Pseudonymisation," adopted for public consultation on 16 January 2025. https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf (accessed 18 October 2026).
- [3] Scale AI, "Data partnerships," web page, undated. <https://scale.com/data-partnership> (accessed 18 October 2026).
- [4] In re Spirit Aviation Holdings, Inc., Bankr. S.D.N.Y., Consumer Privacy Ombudsman supplemental report, Dkt. 1684, 5 October 2026. <https://storage.courtlistener.com/recap/gov.uscourts.nysb.331039/gov.uscourts.nysb.331039.1684.0.pdf> (accessed 18 October 2026).
- [5] Norwegian Computing Center, "Text Anonymization Benchmark (TAB)," corpus repository, MIT licence. <https://github.com/NorskRegnesentral/text-anonymization-benchmark> (accessed 18 October 2026).
- [6] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez and M. Batet, "The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization,"

Computational Linguistics, vol. 48, no. 4, 2022. <https://arxiv.org/abs/2202.00443> (accessed 18 October 2026).

[7] Tonic.ai, "PrivacyBench" (v2.5), dataset and dataset card, CC BY 4.0, revision a79d9e80c5ee68f5dd79da8452e5eb1e1ba97a1e of 15 October 2026. <https://huggingface.co/datasets/TonicAI/Privacy-Bench> (accessed 18 October 2026).

[8] Tonic.ai, "privacy_bench_metrics," scoring code. https://github.com/TonicAI/privacy_bench_metrics (accessed 18 October 2026).

[9] European Data Protection Board, "Guidelines 02/2026 on Anonymisation," draft of 7 July 2026. https://www.edpb.europa.eu/system/files/2026-09/edpb_guidelines_202602_anonymisation_v1_en_0.pdf (accessed 18 October 2026).

[10] S. L. Garfinkel, "De-Identification of Personal Information," NIST Internal Report 8053, National Institute of Standards and Technology, October 2015. <https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf> (accessed 18 October 2026).

[11] U.S. Department of Health and Human Services, Office for Civil Rights, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule," 26 November 2012. <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html> (accessed 18 October 2026).

[12] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz and S. Zanella-Béguelin, "Analyzing Leakage of Personally Identifiable Information in Language Models," IEEE Symposium on Security and Privacy, 2023. <https://arxiv.org/abs/2302.00539> (accessed 18 October 2026).

[13] L. Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Data Privacy Working Paper 3, 2000. <https://dataprivacylab.org/projects/identifiability/paper1.pdf> (accessed 18 October 2026).

[14] J. Branson et al., "Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations," *Trials*, vol. 21, art. 200, 2020. <https://doi.org/10.1186/s13063-020-4120-y> (accessed 18 October 2026).

[15] European Medicines Agency, "External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use," version 1.4, 15 October 2018. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use-version-13_en.pdf (accessed 18 October 2026).

[16] 45 CFR 164.514, "Other requirements relating to uses and disclosures of protected health information." <https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514> (accessed 18 October 2026).

[17] AIData.Broker, "De-identification standard," web page with the engine's catalogue, thresholds and benchmark results, and the in-browser tool at /tools/de-identify. <https://aidata.broker/de-identification> (accessed 18 October 2026).

- [18] Microsoft, "Presidio: data protection and de-identification SDK," MIT licence. <https://github.com/microsoft/presidio> (accessed 18 October 2026).
- [19] Gitleaks, "gitleaks," MIT licence. <https://github.com/gitleaks/gitleaks> (accessed 18 October 2026).
- [20] Internal Revenue Service, "How EINs are assigned and valid EIN prefixes." <https://www.irs.gov/businesses/small-businesses-self-employed/how-eins-are-assigned-and-valid-ein-prefixes> (accessed 18 October 2026).
- [21] US Social Security Administration, top 1,000 baby names by year, 1880 to 2008, public domain, via the mirror at <https://raw.githubusercontent.com/hadley/data-baby-names/master/baby-names.csv> (accessed 18 October 2026).
- [22] US Census Bureau, "Frequently Occurring Surnames from the 2010 Census," public domain. <https://www2.census.gov/topics/genealogy/2010surnames/names.zip> (accessed 18 October 2026).
- [23] Wikidata, given names and family names, CC0 1.0, via the query service. <https://query.wikidata.org/sparql> (accessed 18 October 2026).
- [24] Project Gutenberg, public-domain books (used for word frequencies only). <https://www.gutenberg.org/> (accessed 18 October 2026).
- [25] US Securities and Exchange Commission, `company_tickers.json`, public domain. https://www.sec.gov/files/company_tickers.json (accessed 18 October 2026).
- [26] UK Information Commissioner's Office, "Anonymisation: how do we ensure anonymisation is effective?," 28 March 2025. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/how-do-we-ensure-anonymisation-is-effective/> (accessed 18 October 2026).
- [27] A. Narayanan et al., "On the Feasibility of Internet-Scale Author Identification," IEEE Symposium on Security and Privacy, 2012. <https://intelli-sec.cs.berkeley.edu/papers/sp12.pdf> (accessed 18 October 2026).
- [28] L. Rocher, J. M. Hendrickx and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Communications, vol. 10, art. 3069, 2019. <https://doi.org/10.1038/s41467-019-10933-3> (accessed 18 October 2026).

Statements

Use of AI. This paper was prepared with generative AI tools, which compiled and analysed the open datasets and drafted the text. The author takes responsibility for its content.

Licence: CC BY 4.0. Cite as: Heymans, Y. (2026). De-identifying company workspace exports in the data holder's browser: design and benchmark results. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-deid>