RECOGNICA

Document Processing, on your own server.

Recognica turns the PDFs your business receives every day into verified data. It reads them, recognises what each one is, captures the values you need, and has a person check only what deserves checking. It runs entirely inside your own network.

- OCR & text-layer reading
- Machine-learning classification
- Field extraction
- Human verification
- Windows file shares
- Usage & accuracy reporting
- No external AI
- Unlimited documents, no click charges

Product overview & technology guide
Version 1.0

Recognica, LLC
sales@recognica.net · 813-830-1708
recognica.net

From a pile of PDFs to clean, checked data

Invoices, claims, orders, forms: most organisations still re-key information from documents by hand. Recognica automates that work end to end. Documents are dropped into a shared folder (by people, scanners or other systems) and come out the other side as the original PDF plus a CSV file of the captured values, ready for your accounting, ERP or line-of-business system.

1

Reads any PDF

Uses the PDF's own text when it is there, and OCR when the page is a scan or photo, page by page.

2

Knows what it's looking at

A machine-learning model trained on your own examples sorts each document into its type, with a confidence score.

3

Captures the values

Document Definitions say where each field is found, such as invoice number, date, total or booth number.

4

People check the exceptions

Only documents and fields below your confidence thresholds go to a verifier, who sees the page with each value highlighted.

5

Delivers files you can use

Every finished document is exported as its PDF plus a CSV, in a folder per batch, on a Windows share.

6

Measures everything

Pages, documents, machine time and people's time are recorded for your reporting, along with how accurate the capture was.

Why Recognica

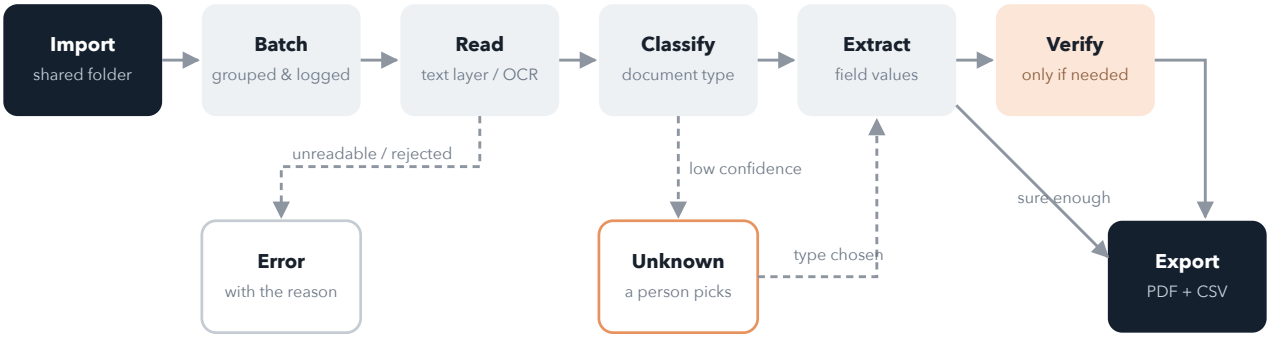
- **Private by design.** Runs on a server you own, inside your network. Documents never leave it.
- **Trained on your documents.** The classifier learns from your own labelled examples, and anything new it meets becomes a training example with one click.
- **People only where needed.** Confidence thresholds send clear-cut documents straight through; verifiers see just the fields flagged "check".
- **Safe changes.** Workflows run on a published version. Edit rules and models freely; nothing changes in production until you publish.
- **Unlimited use, one monthly fee.** No click charges, no per-page or per-document fees: import and export as much as you like. You pay a monthly fee for support and professional services.
- **No external AI.** OCR, classification and extraction run on your server. Recognica makes no calls to any outside AI service or source.
- **Works with what you have.** Scanners and PCs simply save to a network folder; results appear in another one.
- **Keeps the disk tidy.** Retention settings per workflow clean out old files and content automatically, while keeping billing records.

Who it is for

Companies that receive a steady flow of similar documents from many senders: accounts payable, logistics and freight, insurance claims, healthcare administration and back-office services, and service bureaus processing documents for their own clients. Customers, workflows and usage are kept separate, so one installation can serve several clients.

The journey of a document

Each workflow has its own shared folders. The Conductor, Recognica's processing engine, watches them and moves every document through the same steps, timing each one.



1 Documents arrive in Import

Anyone with access saves PDFs into the workflow's **Import** share: a person from Windows or a Mac, a network scanner, or another system. A lock file tells scanners to wait while Recognica is collecting files, so nothing is picked up half-written. System leftovers are cleaned up and stray subfolders are sent to Error with an explanation.

2 A batch is created

Everything collected together becomes a named batch (for example *Invoices-20261113-0930-001*), so it can be tracked, verified and reported as one unit.

3 Pages are read

For each page Recognica decides whether the PDF's own text layer is usable. If it is, it is used as is. If not, the page is rendered and read with OCR. A workflow can also force OCR, or use text layers only. The result is saved, so a document is never read twice.

4 The type is recognised

The classifier scores the document against every type it knows. Above the workflow's classification threshold, the type is set automatically. Below it, the document goes to the **Unknown queue**, where a person picks the type or rejects it, and can add it to the training examples.

5 Fields are captured

The Document Definition for that type finds each field's value and records its page, position and confidence.

6 A person verifies, only when needed

Documents with any field under the field threshold wait for verification. The verifier sees the page with each value boxed, corrects what's wrong, and confirms. Every correction is recorded with who made it and when.

7 Results are exported

Each finished document lands in the **Export** share as its original PDF plus a `.csv` of field names and values, in a folder per batch. Anything that could not be processed goes to **Error**, always with the reason.

Where people come in

Automation does the bulk of the work; people handle the judgement calls. Recognica gives them focused screens and keeps a full record of what they did.

Verification

- Verifiers see the batches assigned to their workflows and reserve one, so two people never work on the same batch.
- Each page is shown with every captured value boxed on it, with zoom and page navigation.
- Fields below the confidence threshold are flagged *check*; the rest show their confidence.
- Keyboard first: **Enter** moves to the next field and **Ctrl+Enter** confirms the document.
- Release a batch for someone else, or cancel to undo everything changed in this sitting.

Unknown queue

- Documents the classifier wasn't sure about wait here with the model's best guess and its confidence.
- A person picks the right type (optionally verifying it right away) or rejects it to Error with a reason.
- Picked documents can be added to the classification project as new training examples, so the model keeps learning.
- Optionally, documents nobody decides on are rejected automatically after a set number of days.

Restarting and reprocessing

- Send chosen documents, a batch, or a filtered set back to any step: OCR, classification, extraction, verification or export.
- Reruns use the workflow's current published rules and model, and can keep values people already verified.
- Each restart is marked billable or not billable, and the choice follows every timing it produces.

Housekeeping, automatically

- Per-workflow retention for working files, Export, Error, and how long field values and detailed history are kept.
- Usage and billing records are always kept, as are accuracy counts.
- Nightly database backups, kept for 14 days.

What comes out

OUTPUT	WHERE	CONTENTS
Data file	Export share, folder per batch	One `.csv` per document: a header row of field names and one row of values, next to the PDF with the same name.
Original PDF	Export share	The document exactly as received, so the image and the data travel together into your systems.
Error files	Error share	Documents that couldn't be read, were rejected, or arrived in the wrong place, each with the reason in its history.
History	Workflows app, Reports	Every step, timing, decision and correction for every document, searchable by file, batch, type, status or person.

Integration made simple: because input and output are ordinary network folders and CSV files, Recognica works alongside almost any ERP, accounting package, RPA tool or scripted import, with no custom connector needed.

Set up a new document process in an afternoon

A **workflow** is one document process for one customer, for example "Vendor invoices for Acme". Administrators build it in the browser with three tools that share one sign-in: **Workflows**, **Document Classification** and **Document Field Extraction**.

1. **Add the customer** Workflows

   Every workflow belongs to a customer, so usage and billing are reported per customer and per workflow.

2. **Teach Recognica your document types** Document Classification

   Create a classification project and drop example PDFs into a folder per type (invoice, credit note, statement…). Recognica reads them, builds the training corpus and trains a model with one click. Each training run is kept in the model history with its training log, so you can compare runs.

3. **Describe the fields** Document Field Extraction

   Create a **Document Definition** per type and upload sample PDFs. For each field you add keyword rules and see the results live on every sample:

   - **Keyword and direction:** the value is to the right of, left of, below or above a label, or on the same line.
   - **Variants:** other wordings of the label, tried in order ("Invoice No", "Invoice #", "Inv. Number").
   - **Value patterns:** what a valid value looks like (date, amount, an ID format), so wrong neighbours are skipped.
   - **Not words:** words that cancel a match on the left or right ("ship" stops "Date" from matching "Ship date").
   - Rename, duplicate and run a whole definition to CSV to check it on all samples at once.

4. **Create the workflow** Workflows

   Name it, pick the customer and the classification project, and choose the settings: classification and field confidence thresholds, how pages are read (automatic, force OCR, or text layer only), how often Import is checked, and retention for each kind of file.

5. **Map document types to definitions**

   Tick the types this workflow accepts and choose the Document Definition for each. Types you don't tick are rejected to Error with the reason.

6. **Choose people and folders**

   Assign verifiers to the workflow and create file-share accounts. Recognica creates the workflow's Import and Export shares, which PCs, Macs and scanners connect to like any Windows share.

7. **Publish and start**

   Publishing freezes a copy of the settings, the rules for each type and the model. The Conductor runs that copy, so later edits never affect running batches. The workflow lists what changed since the last publication. Press Start, and drop the first PDFs into Import.

FEATURES · REPORTING AND ADMINISTRATION

Know exactly what was done, by whom, and how well

Overview

Documents and pages exported today, this week and in any period; what is waiting right now (in the machine steps, to verify, unknown, in Error); batches waiting for verification and who holds them; classification and field confidence per document type.

Usage

For your own reporting or for charging your clients back, per customer, workflow, day and person: documents read, exported and sent to Error; pages split by text layer and OCR; machine time per step; people's active time. Non-billable reruns are left out. Four CSV downloads that add up to the millisecond.

Verification accuracy

For every field, what was captured versus what the verifier confirmed: kept as captured, corrected, cleared, filled in or left empty, with accuracy as a percentage. A list of every correction shows the captured value, the verified value, who and when.

Timing

Average active time per step: per page for OCR, per document and per batch, by workflow. People's time stops after a minute without a key press or mouse movement, so figures reflect real work.

Search and document history

Find any document by file or batch name, type, status or person. Its history shows every event, every field (extracted value, current value, who corrected it), and every timing, with billable flags and reruns.

Server page

Live memory and CPU use per service, disk space and history, with limits that an administrator can change in the browser, including warnings when limits add up to more than the machine has.

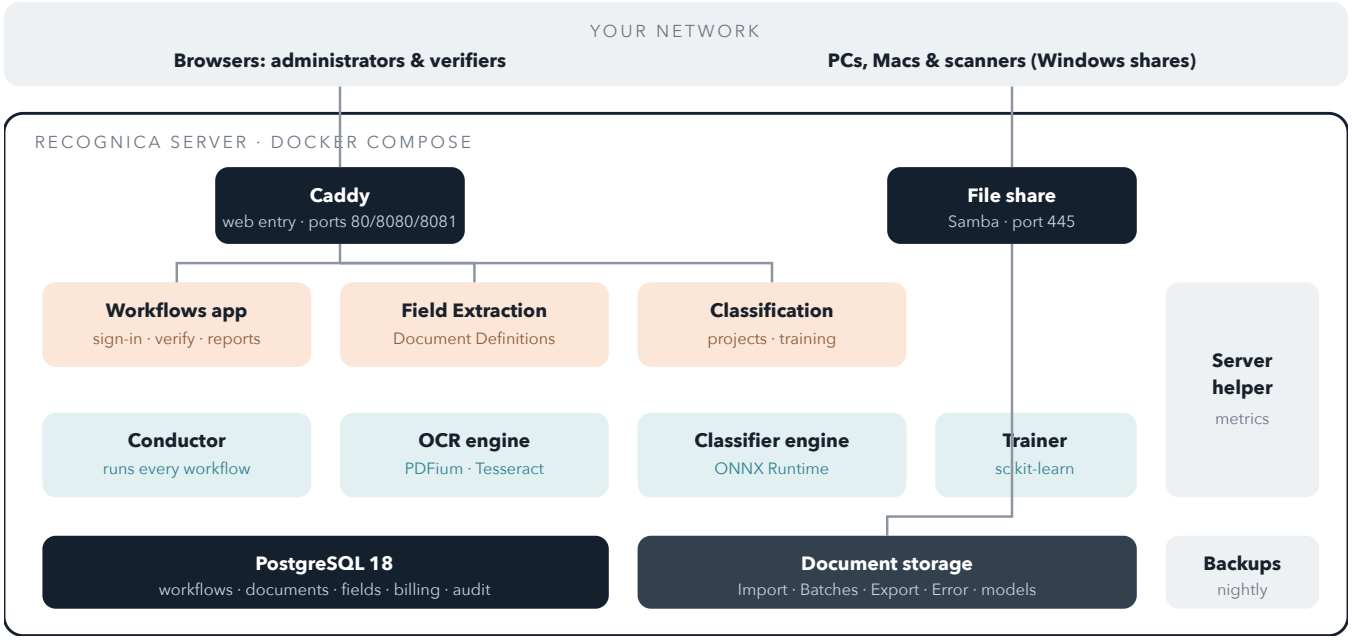
Users, security and audit

- **Two roles.** Administrators build and run workflows; verifiers see only the workflows they are assigned to.
- **One sign-in** for the Workflows app and both design tools; the design tools are for administrators only.
- **Protected accounts:** hashed passwords, lock-out after repeated failures, password changes and resets.
- **First-run setup** with a one-time code from the server's log, so only whoever controls the server can create the first administrator.
- **Append-only records:** timings, events and restarts can't be edited or deleted by the application; corrections are new records.
- **Audit trail** of sign-ins, settings changes, publications, prices, verifications, field edits and file clean-ups.
- **Separate file-share accounts** per workflow, managed in the app.
- **Least privilege inside the server:** each service has its own database login and only the access it needs.

Your data stays yours. Recognica does not send documents, text or values to any outside service. OCR, classification and extraction all run on your server, and processing needs no internet connection.

Architecture

Recognica is a set of small, focused services that run in containers on one Linux server and talk to each other over a private internal network. Only the web entry point and the file shares are reachable from your network.



SERVICE	ROLE
Caddy	Single web entry point. Routes to the three web apps and checks with the Workflows app that a person is signed in as an administrator before opening a design tool. Can serve HTTPS with automatic certificates.
Workflows app	Sign-in, customers, users, workflows and publishing, file-share accounts, verification, the Unknown queue, restarts, reports and the server page.
Conductor	The processing engine. Watches every running workflow's Import, creates batches, calls the engines in order, records timings and events, exports results and applies retention.
OCR engine	Reads text layers, renders pages and runs OCR, page by page in parallel, and times each page for billing.
Classifier engine & trainer	The engine scores documents with a project's model, or a workflow's published copy. The trainer builds new models from the training corpus.
Database, file share, helpers	PostgreSQL holds all records. Samba serves the Import, Export and Error shares. A migration job applies database updates and a backup job dumps the database nightly. The server helper reports resource use and is the only service allowed to adjust container limits.

The technology inside

Reading documents

Text layer	**PdfPig** (.NET) reads the text and word positions already inside digital PDFs. A usability check (characters per page, share of printable characters) sends weak or garbage text layers to OCR instead.
Rendering	**PDFium**, the PDF engine used in Google Chrome (via Docnet), renders scanned pages to images at the resolution OCR needs.
OCR	**Tesseract 5**, the open-source OCR engine originally developed at HP and later Google, using its neural-network (LSTM) recogniser. It returns every word with its position and confidence. Pages of a document are read in parallel.
Saved results	Each page's text, words, positions and confidence are saved next to the PDF, so later steps, restarts and the design tools never have to read the same page again.

Recognising the document type

Model	A text classifier: **TF-IDF** features over words and word pairs (bigrams) feeding a class-balanced **logistic regression**. It is fast, explainable, and accurate on business documents with a modest number of examples per type, with no GPU needed.
Training	Built with **scikit-learn** (Python 3.12) from the text of your labelled examples, with pinned library versions so results are reproducible. Each run is logged and kept in the model history.
Running	Models are exported to **ONNX**, an open standard for machine-learning models, with the vocabulary built in, and run by **Microsoft ONNX Runtime** inside the .NET classifier engine. The same normalised text is used for training and prediction.
Confidence	Every prediction carries a probability per type. The workflow's threshold decides between automatic filing and the Unknown queue.

No external AI. Every model is trained and run on your server. Recognica sends nothing to cloud AI services and makes no calls to any outside source, so there are no per-call fees, no data sharing and no dependency on someone else's service.

Extraction and platform

Capturing fields

Rules engine	Recognica's own keyword-based extraction engine works on the positioned words of each page. It handles labels and values on the same line or across lines, in four directions, with label variants, value patterns (regular expressions) and Not-words to the left or right.
Confidence	Each value carries a confidence from the OCR of its words, compared with the workflow's field threshold to decide whether a person should check it.
Versioning	Rules are saved as JSON with a checksum. Publications store a frozen copy, and every field records which rule version produced it, so reruns can keep verified values where the rule didn't change.

Platform

Applications	**C# on .NET 8** (Microsoft's long-term-support release). The web apps use **ASP.NET Core Blazor Server** for responsive screens without browser plug-ins. Data access uses Npgsql and Dapper.
Database	**PostgreSQL 18**, with versioned migrations applied automatically on start-up, column-level permissions per service, and triggers that keep billing and audit tables append-only.
Web & files	**Caddy 2** web server and reverse proxy. **Samba** for SMB file sharing, compatible with Windows, macOS, Linux and network scanners.
Runtime	**Docker** containers orchestrated with **Docker Compose** on Linux. Each service has its own memory and CPU limits, health checks and automatic restart. It runs on both ARM64 and x86-64 processors.

Built to be dependable

- Each service restarts on its own if it stops, and reports its health.
- Files are only picked up once scanners and senders have finished writing them.
- A batch is reserved by one verifier at a time; administrators can release it.
- The database is updated through numbered, versioned migrations, applied in order.
- Every step's result is saved, so restarts pick up where they left off.
- Billing and audit tables are protected at the database level, not just in the application.

Runs on hardware you choose

Recognica is delivered as one stack of containers for a single Linux server: an on-premises machine, a virtual machine in your data centre, or a compact edge device. The same software runs on ARM64 and x86-64.

SERVER	TYPICAL USE	SPECIFICATION
Edge appliance	Demonstrations, pilots, a single department with light volume	NVIDIA Jetson Orin Nano (ARM64, 6 cores, 8 GB), NVMe storage, Ubuntu-based JetPack Linux. Recognica's reference development system; fits in the palm of a hand.
Standard server or VM	Most production installations	x86-64 Linux, e.g. Ubuntu Server 24.04 LTS · 4+ CPU cores · 16 GB RAM · 100 GB disk plus document storage · fixed IP address.
Recommended	Higher volumes, several customers or workflows, frequent retraining	8+ CPU cores · 32 GB RAM · SSD storage sized to your retention settings. OCR, the busiest step, scales with CPU cores.

No GPU is required: OCR and classification run on the CPU. Memory and CPU limits for each service are set to fit the machine and can be changed from the Server page.

Network

- Web: port 80 (Workflows), 8080 (Field Extraction) and 8081 (Classification), or HTTPS with your own host name.
- File shares: SMB on port 445.
- Everything else stays on the server's private internal network.
- No internet connection, and no outside AI service, is needed to process documents.

Data and backups

- Documents, models and the database are kept in one data folder on the server.
- The database is dumped every night and kept for 14 days, with a ready-made script to copy dumps to another machine.
- Retention settings keep storage use predictable.

Installation

- A single command starts the whole stack; database updates apply themselves.
- A one-time setup code creates the first administrator in the browser.
- A self-contained offline installation package for new servers is in development.

Updates and versions

- Every page shows the running version.
- Workflows keep running on their published version while the software or rules are updated.
- Source and releases are kept in Recognica's own private version control.

PRICING AND PRIVACY

Process as much as you want

Many document-capture products charge per page, per document or per "click", so your bill grows with your success. Recognica doesn't.

No click charges

No fees per page, per document or per field. Volume never changes the price.

Unlimited use

Unlimited importing, exporting, workflows and reprocessing, limited only by your server.

One monthly fee

A predictable monthly fee for support and professional services: set-up, new workflows, tuning and updates.

100% internal: no external AI

Nothing leaves your network

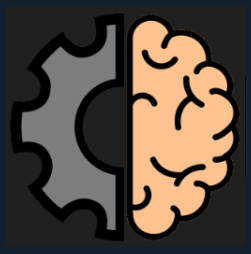
OCR, classification, field extraction, training and reporting all run on your own server. Documents, text and values are never sent to a cloud service.

No outside calls

Recognica makes no calls to external AI services or any other outside source while it works, so there's no third-party data processing to approve and no surprise usage bills.

Feature summary

- Watched Import folders per workflow, scanner-safe
- Automatic text-layer detection with OCR fallback
- Force-OCR and text-layer-only modes
- ML document classification with confidence
- Unknown queue with one-click training examples
- Keyword, variant, pattern and Not-word field rules
- Live rule testing on sample documents
- Verification with highlighted values, keyboard-driven
- Batch reservation, release and cancel
- PDF + CSV export per document
- Publish and version workflows safely
- Restart from any step, billable or not
- Usage, accuracy and timing reports with CSV
- Document search and full history
- Customers, roles, verifier assignment
- Per-workflow retention and cleanup
- Append-only billing and audit records
- Server monitoring and limits in the browser



Let's look at your documents.

The fastest way to see what Recognica can do is with your own paperwork. Send us a handful of sample PDFs and we'll show you them classified, captured and exported, on a server that never leaves your control.

Unlimited documents · No click charges · No external AI · One monthly fee

recognica.net

sales@recognica.net · 813-830-1708

Recognica® Document Processing, version 1.0 · © 2026 Recognica, LLC. All rights reserved.
Product names mentioned (NVIDIA, Jetson, Microsoft, .NET, PostgreSQL, Tesseract, Docker and others) are trademarks of their respective owners.