

Classifix.

Turn unstructured documents into categorised, machine-readable data — without sending them anywhere.

Upload documents, read them with on-premise OCR in 129 languages, then classify and extract fields using language models that can run entirely on your own hardware. People verify the results in a keyboard-driven review station with the recognised words drawn straight onto the page.

129 OCR languages

9 model providers

Air-gappable

236 REST endpoints

Self-hosted

129

BUNDLED OCR LANGUAGES

3

OCR OUTPUT FORMATS

17

TRACKED JOB TYPES

0

BYTES THAT MUST LEAVE

THE PROBLEM

The documents you most need to process are the ones you cannot send away.

Cloud document-understanding services are excellent, and for a large class of buyer they are simply not an option: the documents are patient records, case files, tender responses or personnel data, and the moment they cross a border or a tenancy boundary the project stops. So the work stays manual — sorting, keying, filing — at a cost nobody measures because it has always been that way.

Data cannot leave

Regulation, contract or policy rules out any service that processes documents off-site.

Mixed inbound batches

A scan run containing five document types, and someone paid to split it by hand.

Keying from paper

Fields retyped into a system that could have read them, with the errors that implies.

No feedback loop

Corrections vanish into a person's head instead of improving the next batch.

THE SOLUTION

The entire pipeline runs on your hardware — including the model.

Local recognition

OCR runs in-process with 129 language packs already in the product. No download step, no per-page service, no outbound call.

Local intelligence

Nine model providers behind one configuration — five of them run on your own machines. Classification and extraction with zero data egress is a setting, not a project.

People in the loop

A review station built for speed: the document on one side, the decision on the other, and the whole queue navigable from the keyboard.

THE PIPELINE

1

Ingest

Single or bulk upload by category, drag-and-drop, or through the API. Large files arrive in chunks. Storage on local disk or a connected cloud folder.

2

Read

Pages rasterised at 300 dots per inch and recognised, producing plain text, word-positioned markup and an archival XML layer — all three kept, all three reusable.

3

Understand

Classify into your own categories and extract the fields you define, using a prompt you control against a model you choose.

4

Verify

Results queue into the validation station, where a reviewer accepts or re-files each document against the page image with the recognised words overlaid.

READ & UNDERSTAND

Recognition, classification and extraction.

1 Optical character recognition

- **129 language packs bundled in the product** — including vertical Chinese, Japanese and Korean, Fraktur and script detection
- Three outputs kept per document: plain text, **word-positioned markup** and an archival XML layer
- PDF pages rasterised at 300 dots per inch; images accepted directly
- Run on one document, a whole category, an entire classifier, a single page, or an ad-hoc upload
- Thread pool sized to the machine; per-page progress; results cached so re-running is a choice, not a cost
- Configurable resolution, language, page-segmentation mode and output formats
- Encrypted and zero-page PDFs detected and rejected rather than silently producing nothing

2 Classification

- Define your own classifiers and your own categories, with bulk category creation from a pasted list
- **Classification by language model** — returns the category, a confidence and its reasoning
- **Per-page classification with automatic splitting**: a mixed inbound batch is separated into one output file per category
- A verification mode that has the model check a classical classifier's answer
- **Category keywords** — required, optional and prohibited terms that score a document and can re-file or flag it automatically
- Classical text classifiers available alongside, with production inference on two of the four bundled frameworks
- Reclassification workflow: re-run over training documents, review mismatches, apply one at a time or in bulk

3 Field extraction

- Extractors defined per document type with four prompt slots — system, extraction, schema and refinement
- An optional **second refinement pass** over the model's own first answer
- Output as JSON or XML, stored with the field count, confidence, method, model and processing time
- Single, sequential batch and parallel batch extraction endpoints
- A regex path for simple, fixed patterns where a model is unnecessary

4 Models, local or hosted

- **Nine providers** — five that run on your own hardware and four hosted
- **Named model profiles** per organisation with full generation-parameter control, and a connectivity test
- Reusable, versioned prompts with **templated variables** for both classification and extraction
- A model pass that corrects recognition output before anything downstream reads it

The word overlay is the quiet differentiator. Because recognition keeps word positions, the viewer draws every recognised word box straight onto the page. A reviewer sees exactly what the machine read and where — which is the difference between checking a result and merely trusting one.

VERIFY, OPERATE & INTEGRATE

Around the pipeline.

5

The validation station

- A three-panel review screen: queue, document, decision
- **Keyboard-driven** — next, previous, accept and page navigation without touching the mouse, and suppressed while typing
- The page image with recognised word, line and block boxes overlaid
- Accept a classification or re-file it against the category list, one document at a time
- Queue statistics so a supervisor can see the backlog rather than guess at it

7

Jobs & operations

- **Seventeen tracked job types** covering recognition, training, classification, extraction, reclassification, restore and housekeeping
- Every job reports progress, items processed, the item in flight, an estimated finish and a detailed log
- Cancel a job, re-run a failed one, bulk-clear finished ones; administrators see every organisation's jobs
- Background work keeps its organisation scope through a dedicated system identity
- Log console with wildcard search, level filtering, tailing and download; memory and storage housekeeping with a manual trigger

6

Evaluation & comparison

- A **model bake-off** that runs a set of named models over the same corpus and reports accuracy, average confidence and processing time
- Per-category precision, recall and F-measure with the underlying counts
- Test runs persisted as history, so two runs can be compared on a chart rather than from memory
- Model backup and restore with timestamped sets

8

Integrate & move

- **236 REST endpoints** across twenty-three controllers, described by an OpenAPI document
- **Scoped API keys** with a per-key permission map, expiry, revocation and per-day usage accounting
- Export a whole classifier — files and data — as an archive, and import it elsewhere
- **Stream a classifier directly from one Classifix instance to another**, or copy it within one
- Guided first-run setup, including generation of a service unit
- Seventeen in-product help articles with full-text search

Where the product is today. Language-model classification and extraction are the production path and are what we recommend building on. Classical machine-learning classification runs inference on two of the four bundled frameworks; the others are training and evaluation only. Rate limiting, encryption at rest and single sign-on are not in this release. We would rather you had that from us than from a pilot.

WHO BUYS IT

Document volume plus a reason it cannot leave the building.

Accounts payable

Invoices and receipts sorted by supplier and type, with header fields lifted straight off the page instead of keyed.

Accounting & bookkeeping firms

Client documents arriving in mixed batches, separated automatically and filed per client under isolated classifiers.

Legal

Case bundles and disclosure sets, sorted by document type, where sending the material to a hosted service is not permitted.

Healthcare

Referrals, reports and forms processed inside the clinical network, with a human accepting every classification.

Public sector

Correspondence and application handling in environments where an air-gapped deployment is the only deployment.

Defence & regulated industry

The whole stack — recognition, model, database — runs with no outbound connection at all.

Document bureaux & BPO

Per-client classifiers, isolated categories, and the ability to move a trained classifier between installations as a deliverable.

Insurance

Claim packs split per page into their constituent document types before anyone opens them.

Logistics & customs

Mixed shipping paperwork separated and read, with the recognised text kept for search and audit.

Human resources

Personnel files and applications categorised inside the organisation's own boundary.

Cross-border operations

129 recognition languages in the box, so a multilingual mailroom is a configuration rather than a procurement.

Software vendors

A documented API and scoped keys for embedding classification and extraction inside your own product.

WHO USES IT

Platform operator

Organisation administrator

Document worker

Validation reviewer

Read-only viewer

API client

Four roles plus two non-human identities: a scoped API client with its own permission map and usage accounting, and a system identity that carries organisation scope into background work.

AGAINST THE ALTERNATIVES

OPTION	WHY IT FAILS HERE
Hosted document AI	Documents leave the boundary; priced per page
Manual sorting and keying	Scales only by hiring, and the errors are invisible
A scanning tool with templates	Breaks the moment a supplier changes their layout
A bespoke script	No review station, no queue, no audit, one maintainer

TECHNICAL SPECIFICATION

One JAR, one database, and no outbound dependency.

Runtime	Java 21 on Spring Boot, single self-contained JAR with the recognition engine and language packs included
Database	PostgreSQL
Front end	Server-rendered with vanilla JavaScript — no framework, no build step
Viewer	In-browser PDF rendering with a word-position overlay driven by the stored recognition layer
Storage	Local disk or a connected cloud folder, in a per-organisation, per-classifier tree, with chunked upload for large files
API	236 endpoints across 23 controllers with an OpenAPI document
Models	Nine providers, five of them self-hosted
Install	Guided first-run wizard including service-unit generation
Scale	~139 000 lines across application and tests; 24 entities; 168 test classes

SECURITY POSTURE, STATED PLAINLY

What this release implements:

- Four roles enforced through method-level authorisation across the API
- bcrypt password hashing; stateless signed session tokens
- Scoped API keys with a permission map, expiry, revocation and daily usage accounting
- Created and modified stamps plus optimistic locking on every record
- An explicit cross-origin allow-list and frame protection
- Dependency vulnerability scanning available in the build

Not in this release: multi-factor authentication, single sign-on, rate limiting, encryption at rest, self-service password reset, and data-subject export or erasure tooling. Organisation separation is by scoping in each service rather than enforced in the database. Deploy on a trusted network behind your own access layer, and ask for the current roadmap before positioning against a compliance requirement.

COMMERCIAL

Licensed per installation. No page meter.

The hosted alternatives charge per page, so success costs more. Classifix is bought once and run by you, and with local recognition and a local model the marginal cost of a page is the electricity to process it.

REVENUE	PER YEAR	REVENUE	PER YEAR
Up to €500k	€900	€100M	€13,000
€1M	€1,500	€500M	€37,000
€5M	€2,500	€1B	€52,000
€25M	€5,500		

Annual licence on your own turnover, ex VAT. The schedule is marginal like income tax — each slice of turnover is charged at its own rate, so there are no cliffs — with a €900 minimum and any yearly increase capped at 25%. Every tranche carries unlimited users, projects and organisations inside one tenant, self-hosting and support that rises with the tranche. Hosting and hardware are yours, AI runs on your own provider keys and is billed by them, and consultancy, setup, migration and training are quoted separately. Illustrative: a metered capture service at €0.02 a page bills a two-million-page year at €40,000; at €5M turnover this is €2,500.

NEXT STEP

Send us a real batch.

A mixed scan run, as it actually arrives. We will split it per page, classify it and show you the review queue.

Sales	tim@automatize.eu
Partners	luc@automatize.eu
Web	automatize.eu