

Kognit Capture.

Capture, extract and structure data from any document — on your own infrastructure, with no page meter running.

An API-first intelligent document capture platform. Documents arrive from thirteen channels, pass through a sixteen-step image pipeline, are read by local OCR, local models or twelve remote providers, are mapped onto versioned templates, reviewed by people where it matters, and delivered to thirteen kinds of downstream system. Built as a modern replacement for legacy capture suites.

Self-hosted

Multi-tenant

129 OCR languages

13 LLM providers

No licence gate

632

REST ENDPOINTS

31

FIELD TYPES

109

AGENT TOOLS

€0

MARGINAL COST PER PAGE

THE PROBLEM

Document capture is priced by the page and delivered by the quarter.

The incumbent platforms in this category share two properties. They meter you per page, so the more value you extract the more you pay. And they take eight to twelve weeks of specialist consulting to configure a single document type. Meanwhile the cloud alternatives require your invoices, claim forms and identity documents to leave your building — which for a growing number of buyers ends the conversation.

Page meters

Volume growth becomes a cost problem instead of an efficiency gain.

Forced migrations

Vendors retire the platform you built on and hand you a rebuild project.

Data egress

Cloud capture means sending regulated documents to somebody else's tenancy.

Consulting dependency

Every new document type is a change request, not a configuration.

THE SOLUTION

One platform, on your hardware, with the economics inverted.

No runtime licence gate

There is no licence server, no entitlement check, no page counter. Run local OCR and local models and the marginal cost of a page is the electricity to process it.

Blueprints, not projects

Six industry starter kits provision an entire working project — workflow, steps, routing, upload endpoint, export goal and a fully fielded template — in one transaction.

Nothing leaves the building

Tesseract and local models run in-process. Air-gapped operation is a supported deployment, not a workaround. Remote model providers are an option, never a requirement.

THE PIPELINE

1 Ingest

Upload, REST, hot folder, SFTP/FTP, e-mail, message bus, browser scanning hub or a paired handset — thirteen import channels in all.

2 Prepare

Sixteen image operations: deskew, despeckle, binarise, contrast, sharpen, crop, line and punch-hole removal, colour dropout, background flattening and more.

3 Read

Eight recognition modes across Tesseract, native PDF text, local models and remote providers, with grammar-constrained structured extraction.

4 Verify

Operators correct results in browser hubs with the page image side by side, driven by validation rules and confidence — with a full field-level trail.

5 Deliver

Thirteen export destinations and eight formats, including searchable PDF and PDF/A built from persisted word geometry.

CAPTURE & RECOGNITION

Getting the document in, and reading it correctly.

1 Ingestion & scanning

- Thirteen import types: upload, REST, SFTP, FTP, hot folder, e-mail, Kafka, RabbitMQ, ZeroMQ, webhook, script, plugin, dataset
- **Driverless network scanning** — server-side eSCL discovery, or direct-attach through the browser. Nothing to install on the scanning desk
- Scan Hub sessions that survive a closed laptop: reorder, rotate, delete, rescan, separate, then submit as one batch
- Archive expansion, PDF-and-XML pair handling, embedded e-invoice payload extraction
- Paired mobile capture with revocable per-device keys
- Scheduled polling with logged results per source

3 Templates & extraction

- **31 field types** and **10 extraction methods** — zone, label search, regex, keyword proximity, model, manual, anchor text, companion file, classification, XML mapping
- Versioned templates that integrations pin to, so a design change never breaks a live interface
- Table and line-item extraction with column zones, row grouping, header detection and continuation across pages
- Template alternatives — layout variants per supplier, merged by field name
- Template generation from a single sample page, and training from operator corrections
- Reference datasets for lookup, autofill and validation

2 Recognition

- Eight engine modes across Tesseract 5, native PDF text, hybrid and open data loading
- **129 bundled language packs**, warm pooled engines, no external OCR service
- Local models running in-process, plus twelve remote providers behind one interface with failover
- Grammar-constrained extraction built from the template's own field set, including per-row line items
- Barcode and QR reading on a dedicated decode path — never guessed from OCR text
- Word-level boxes and confidence persisted, which is what makes searchable PDF and PDF/A output exact
- Output as hOCR, ALTO XML or plain text

4 Classification & splitting

- Four machine-learning frameworks and **29 model types** running in-process — no external inference service
- Classification by model, by language model, by rules, or hybrid; with test runs and prediction review
- Document splitting on fixed page count, blank page, barcode match, model detection, template header/footer or physical separator sheets
- Parent-child split lineage retained

Six blueprints, working on day one. Finance (bank statements), Accounts Payable (supplier invoices with line items), Insurance (claims), KYC onboarding (identity documents including the machine-readable zone), Anti-Money Laundering (transaction reports) and Supply Chain (bills of lading). Each provisions a complete project in a single transaction — against an industry norm of eight to twelve weeks of implementation.

WORKFLOW, PEOPLE & DELIVERY

Where the document becomes trusted data.

5 Workflow engine

- **16 step types:** import, OCR, model OCR, classify, recognition, verification by model, processing hub, verification hub, template training, export, processed, custom script, companion data, XML mapping, cross-check, plugin
- A visual designer where the drawing *is* the workflow — decision gateways on document properties and field values
- Workflow versioning, validation, publish, backup and restore
- **14 validation rule types** from not-empty and ranges through checksum, expression, script, REST call, database query and dataset lookup
- Custom scripts in JavaScript, Groovy or Java, sandboxed, versioned and testable against sample documents
- Service-level monitoring, stalled-work sweeping and document priority

6 Human review

- Three operator hubs — **Verification, Processing and Scan**
- Page image beside the fields, click-to-fill from the image, zone overlay, keyboard-driven review
- Per-field trail: raw recognised value, original value, who verified it and when
- Check-out locking so two operators never touch the same work
- Public single-document review links for an outside party with no account

7 Delivery & integration

- Thirteen export destinations, eight formats including searchable PDF and **PDF/A-2b**
- Field mapping per export goal — unmapped fields are dropped by design
- Eleven external connection types with a wizard and a test action for each
- Encrypted per-project credential vault
- Plugin SDK for compiled workflow steps, importers and exporters — with their own migrations and endpoints
- **Embeddable hubs:** put the Verification or Scan Hub inside your own product in an iframe, behind an API key, a short-lived session and a named origin

8 Built for agents

- A native **Model Context Protocol server** with around 109 curated tools across projects, documents, verification, recognition, templates, workflows and administration
- OAuth 2.1 with dynamic client registration and an explicit consent screen
- Scoped tools with risk tiers, and a separate audit channel that records agent *reads*, not only writes

Scale out

Duty- and capability-aware cluster nodes claim work without contention. Drain a node, resume it, or dedicate it to OCR, export or import.

Documented in the product

361 help articles across 33 chapters, plus an OpenAPI console — permission matrix, step matrix, storage layout, schema reference and error index included.

Honest about limits

The repository ships its own competitive analysis, including where local OCR loses to commercial engines on degraded scans. We would rather you knew.

WHO BUYS IT

Where documents are the raw material of the business.

Accounts payable

Supplier invoices with line items, per-supplier layout learning, three-way match data and ledger export. The highest-volume, fastest-payback use case in the catalogue.

Banking & finance

Bank statements with transaction line items, account reconciliation feeds, and document sets that cannot be sent to a public cloud.

Insurance

Claim forms and policy schedules with damaged-item and coverage tables, routed to adjusters through review hubs.

KYC & onboarding

Identity documents down to the machine-readable zone, proof of address, with one-hundred-percent human review where regulation demands it.

Anti-money laundering

Transaction and suspicious-activity reports with validation rules and a regulatory filing export.

Logistics & customs

Bills of lading, packing lists and customs declarations with reference numbers, economic operator identifiers and container tables.

Capture bureaux & BPO

Twenty client companies on one installation, isolated down to the filesystem. This is the strongest single wedge against per-install legacy suites.

ISVs & OEMs

Embed the Verification and Scan Hubs inside your own application and sell capture as part of your product, not as a partner referral.

Public sector & archives

Searchable PDF and PDF/A renditions, 129 languages, and a documented air-gapped deployment for material that cannot leave the network.

E-invoicing recipients

Intake of Peppol, UN/CEFACT CII, ZUGFeRD, Factur-X and XRechnung documents alongside scanned paper, mapped onto the same template.

Shared service centres

One platform, many business units, each with its own projects, templates, operators and exports — under one operations team.

Sovereignty-bound buyers

Defence-adjacent, healthcare and government organisations for whom the deciding question is simply where the document is processed.

WHO USES IT

Platform operator

Organisation administrator

Verification operator

Mailroom / scanning clerk

Capture engineer

Solution architect

Script & plugin developer

API integrator

OEM integrator

External reviewer

AI agent

Roles are granted per project as well as globally, and the published permission matrix ships inside the product.

THE DISPLACEMENT MOTION

INCUMBENT SHAPE	THE OPENING
Legacy capture suite facing end of life	A forced rebuild anyway — better spent on a platform that will not meter you
Per-page cloud extraction	Cost scales with success; data leaves the tenancy
Per-install desktop capture	No tenancy, no API surface, no agent access
Custom in-house scripts	No review hub, no audit, no versioning, one person who understands it

TECHNICAL SPECIFICATION

One JAR, one database, your hardware.

Runtime	Java 25 on Spring Boot 4, single self-contained JAR with OCR libraries bundled; optional native image
Database	PostgreSQL, 90 versioned migrations, every table and column commented in-database
Front end	Server-rendered with vanilla ES6 — no framework, and no remote asset, enforced by a test
Storage	Files never in the database. Tenant-prefixed disk tree; the database holds paths only
API	632 endpoints across 83 controllers, OpenAPI 3 with a built-in console
Agents	Model Context Protocol server over streamable HTTP, OAuth 2.1
Scale	Duty- and capability-aware cluster nodes, contention-free job claiming, drain and resume
Extensibility	Compiled plugin SDK plus sandboxed scripts in three languages
Codebase	~160 000 lines, 90 entities, 384 test classes, 361 help articles

SECURITY & GOVERNANCE

- Role-based access with per-project grants and a published permission matrix
- Tenant isolation at data *and* filesystem level, through a single path service
- **AES-256-GCM encryption at rest** for provider keys and project credentials; API keys stored hashed
- Central audit log of logins, creates, updates, deletes, processing, verification, import and export — plus a separate agent audit channel
- Login throttling and account lockout
- Script sandbox with no filesystem, network or cross-tenant access
- Origin allow-lists and access policy for embedded hubs and API keys
- Per-project retention policies with soft delete, file deletion or cascading hard delete
- A written compliance chapter covering GDPR, HIPAA and CCPA — as a gap analysis, not a claim

Self-hosting places certification obligations on the operator. KognitCapture carries no vendor SOC 2 or ISO 27001 attestation, and the product documentation says so plainly. Multi-factor and single sign-on are on the roadmap, not shipped.

COMMERCIAL

Licensed per installation. Never per page.

Buy once for the platform and run the volume you have. The software cost is fixed and predictable, which keeps your budget where the value is — the templates, the integrations and the review process around them.

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 fifty of your worst documents.

Not the clean ones. We will build the template, run them through, and show you the extraction and the review effort honestly.

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