AI-Powered Document Extraction for Pharmaceutical Compliance

1. The Compliance Imperative

- **Unstructured avalanche** – ≈80 % of enterprise information now sits in PDFs, legacy scans and other unstructured formats that cannot be queried directly.

- **Regulation is becoming data-centric** – EMA's SPOR/PMS programme is rolling out ISO IDMP submission in phases; structured product data is already live and will expand in 2025–26.

- **Manual extraction is brittle** – human first-pass accuracy rarely exceeds 90%; throughput is constrained to minutes per page; expensive subject matter expert costs scale linearly with volume.

- **Business risk** – delayed or inconsistent data can hold up variations, trigger deficiency letters, lead to re-work across markets, or risk significant penalties.

2. Document Processing Solution

The initiative replaces slow, repetitive reading and copy-paste work with an **automated pipeline** that ingests legacy documents visually, recognises layout, extracts required fields, and delivers *ready-to-submit* IDMP JSON to a RIM.

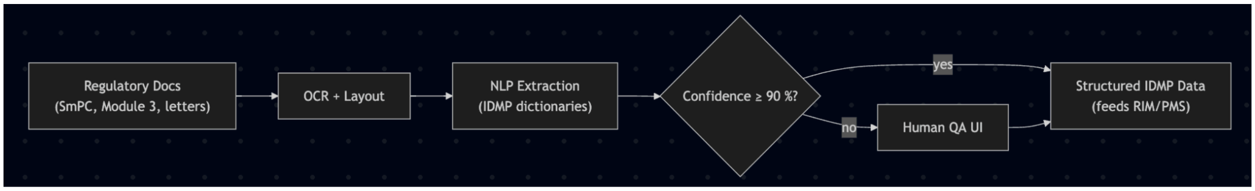

Figure 1 – automated pipeline with optional human review.

Key business attributes

Attribute	Why it matters	How we addressed it
Accuracy	Regulators expect data fidelity	Model tuned to pharma language / structured vocabularies; human review on low-confidence fields → **>95 % field-level accuracy**

Speed	IDMP deadlines & rolling variations	<1s / page vs. ~5 min manually → >300× faster
Cost	Control SG&A and avoid overflow headcount	McKinsey benchmarks show ≥ 30 % cost savings on data management when traditional intelligent document processing is adopted in life-sciences: our solution saves closer to 90%.
Security	Patient & product data subject to GDPR/Part 11	Choice of three deployment tiers (below)

3. Deployment & Security Tiers

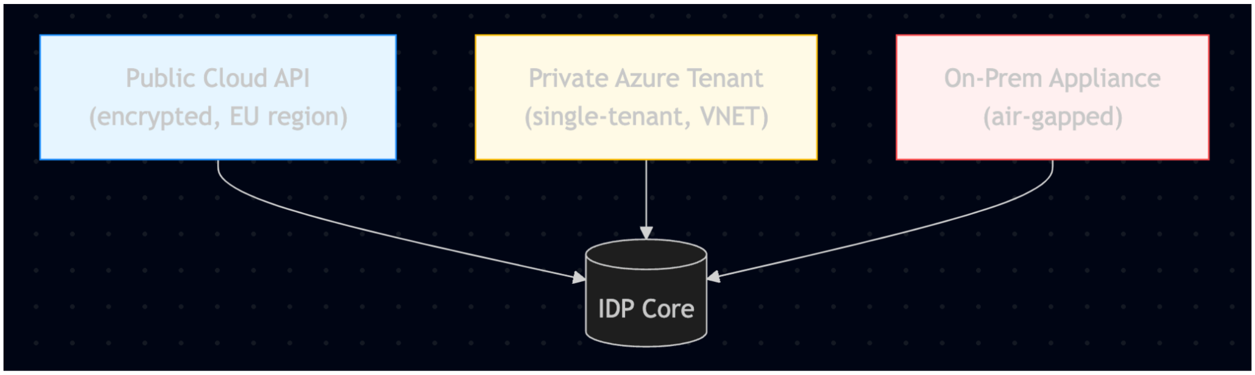

Figure 2 – deployment choices mapped to data-sensitivity for Intelligent Document Processing (IDP).

- **Cloud API** – fastest start-up, suitable for low-sensitivity docs.
- **Private Azure** – used for main rollout; meets GDPR residency with full audit trail.
- **On-Prem** – for documents that cannot leave network (e.g., patient identifiers).

All tiers encrypt data in transit / at rest, integrate with SSO & MFA, and log every action for FDA 21 CFR 11 audit readiness. Pipeline is extremely lightweight with minimal dependencies (Minimal OS with simple file system, Python runtime).

4. Results (2-month roll-out)

KPI	Before (manual)	After (IDP)	Δ
Avg. extraction accuracy	~90 %	**96–97 %**	+7 pp
Pages processed / FTE-day	~200	**20 000+**	100×
Regulatory staff hours saved	—	**≈ 8 000 h** (IDMP wave 1)	—
First-year ROI	—	**> 300 %** (labour avoided vs. licence + cloud)	—

(Internal audit sample; figures rounded.)

Business narrative: Cleared IDMP backlog in record time, mitigating penalty risk and substantial labour costs, without the expense of accuracy.

5. Roadmap & Strategic Value

- **Continuous IDMP iterations** – new EMA data fields can be added by configuration, not new projects.
- **Reuse across functions** – same pipeline can harvest data for pharmacovigilance, quality, clinical reports.

6. Executive Take-aways

1. **Automate where humans add least value** – reading thousands of complex static documents is prime territory.
2. **Start with a clear compliance target** – IDMP provided a forcing function and measurable win.

3. **Architect for trust** – selective human QA plus strong audit trails convert sceptics.

Intelligent document processing is now a board-level lever: it cuts cost, accelerates submissions, and lays the data foundation for next-generation regulatory operations.

Sources:

https://deep-talk.medium.com/80-of-the-worlds-data-is-unstructured-7278e2ba6b73

https://www.ema.europa.eu/en/human-regulatory-overview/research-development/data-medicines-iso-idmp-standards-overview/substance-product-organisation-referential-spor-master-data/substance-product-data-management-services

https://emerj.com/intelligent-document-processing-financial-services-two-use-cases/

https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality