Atlas OCR

An Experimental Concept for Turning Scanned Documents into Accessible PDFs

Making document accessibility a computational problem instead of a labor problem.

Atlas Systems · Experimental Concept

The Challenge

Today's tools make you choose: OCR or accessibility markup. Not both.

OCR Tools

Extract text and make PDFs searchable — but produce no accessibility markup. No heading tags, no reading order, no screen reader support.

Accessibility Tools

Add tags and structure to existing documents — but don't handle scanned images or perform text extraction. They need text to already be there.

Neither Handles Language

Most tools assume English. They don't detect languages from the image or apply the right processing for Arabic, Japanese, or mixed-language pages.

Atlas OCR combines OCR, accessibility markup, and automatic language detection into a single pipeline — so libraries don't have to piece together separate tools.

Built for Our Workflows

This experimental project didn't start as a standalone product idea — it came directly from a customer need.

Where It Started

Ares customers told us they need to meet ADA Title II accessibility requirements for course reserves. Articles and documents served to students through Ares must be accessible — and right now, most of them aren't. That's where this started.

What We're Building

A workflow that extracts course articles from Ares, runs them through accessible markup processing, and pushes the compliant documents back into the system — automatically. No manual remediation, no separate tools, no extra steps for library staff.

The Bigger Picture

As part of addressing the accessibility markup need, we're researching and developing this experimental OCR concept to meet those requirements — and it's designed to integrate across our product suite, serving the same compliance needs wherever documents are delivered.

Atlas OCR: The Solution

An experimental pipeline that combines OCR, accessibility markup, and language detection into a single tool — purpose-built for library workflows.

Accessibility-First

Produces tagged PDFs with heading hierarchy, reading order, tables, and language markup — not just searchable text.

Multi-Model Architecture

Orchestrates multiple specialized AI models together — each doing what it's best at — so the pipeline is smarter than any one model alone.

Purpose-Built for Libraries

Designed to plug directly into the workflows librarians already use — like Ares — for both batch backlog processing and daily inline workflows.

Proven on Hard Documents

Validated on Arabic RTL, Japanese vertical text, handwritten tables, degraded mimeographs, and multi-column layouts.

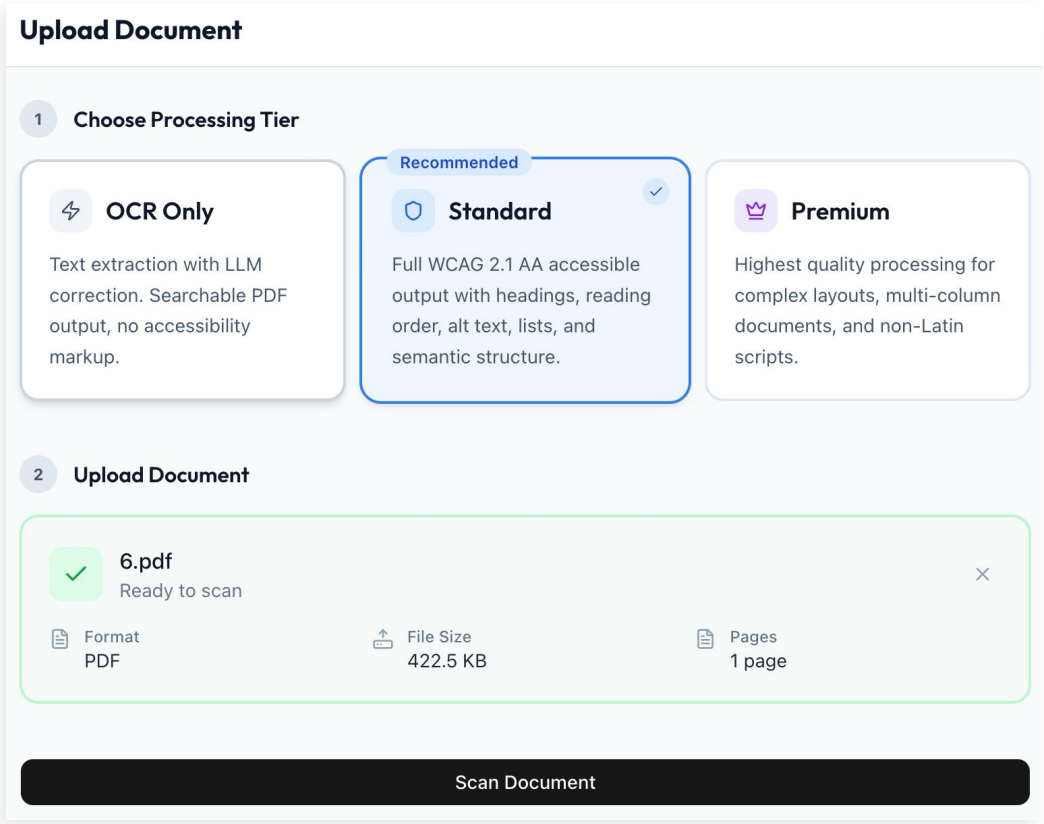
How It Works

The system uses AI to understand a page's visual structure — headings, paragraphs, tables, images, reading order — then extracts and corrects the text using processing tailored to what it found, and assembles everything into an accessible tagged PDF that a screen reader can navigate.

The key insight is that no single AI model is good at all of these tasks, so we orchestrate multiple specialized models together, and the pipeline is smarter than any one model alone.

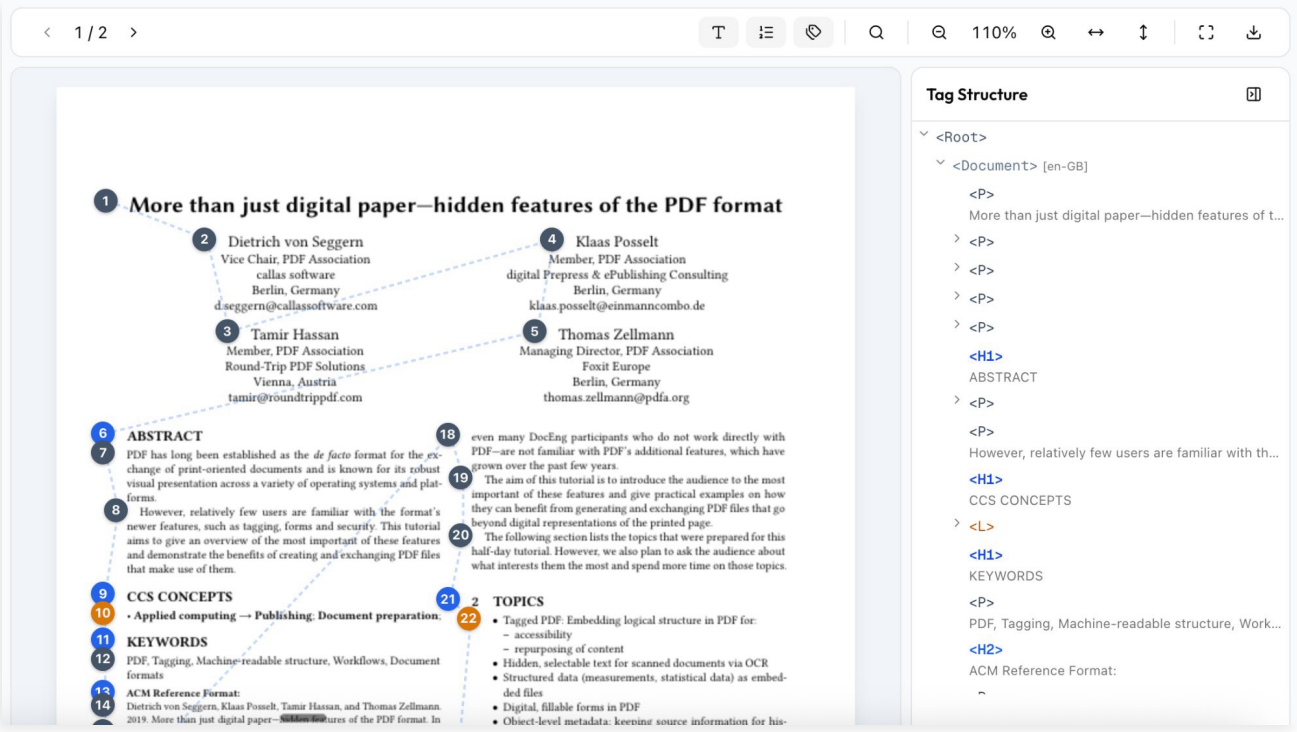
Simple to Use

Library staff choose a processing tier, upload the document, and click Scan. No technical expertise required.



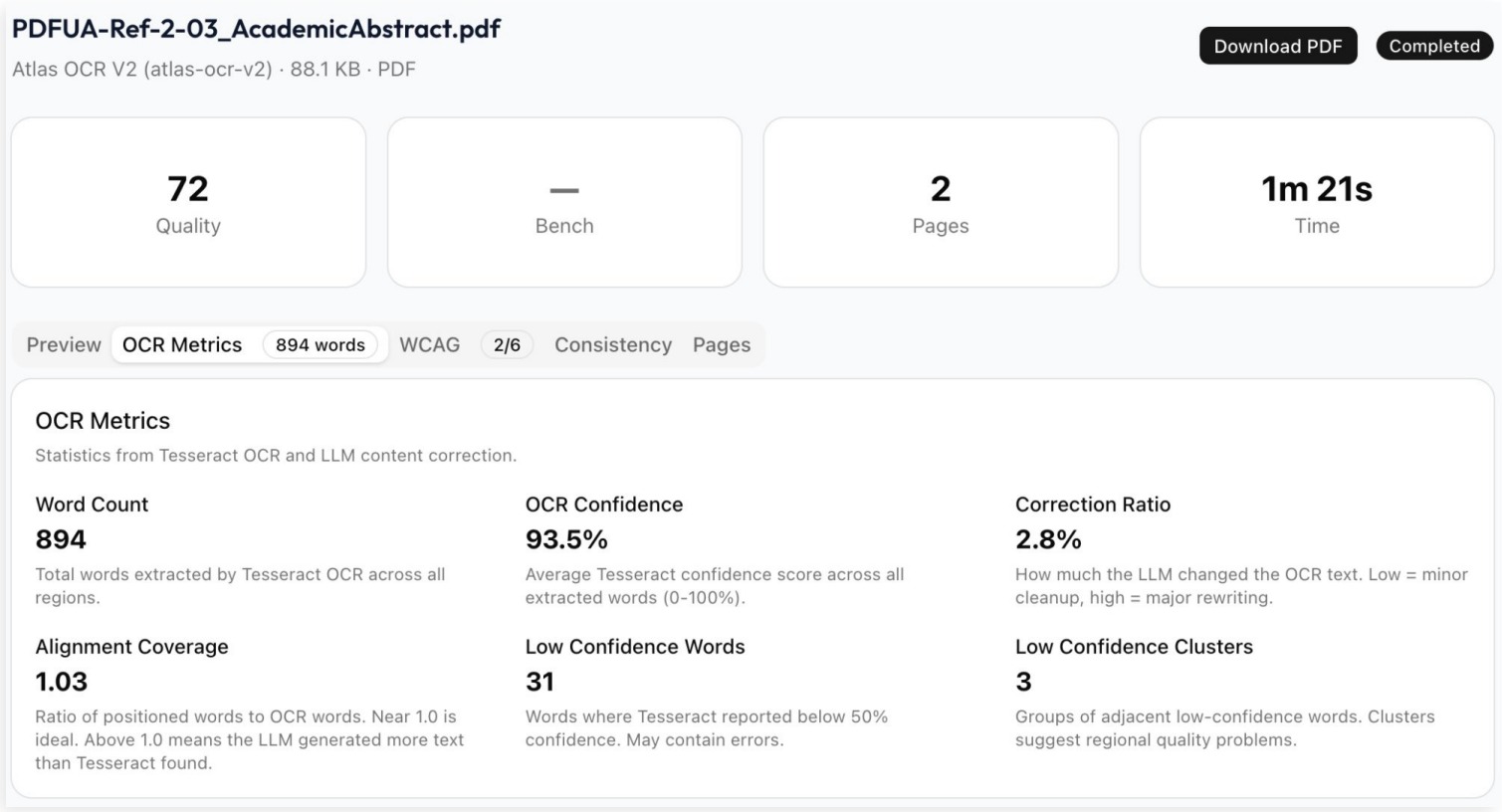
Built-in PDF Visualization

A tool built into the system lets you inspect the output PDF's tag structure, reading order, and region detection — so you can see exactly what the pipeline found.



Quality at a Glance

Every processed document gets a detailed quality report — word count, OCR confidence, correction ratio, and WCAG compliance results — all surfaced automatically.



How We Validate

We're actively running these tests to guide development — the pipeline is still being refined, and we're working through edge cases to meet these marks consistently.

1

WCAG Compliance Validation

Output PDFs are validated against the full WCAG 2.2 profile using veraPDF — the same tool used by the PDF/UA standards body. Automated rules checked on every single page.

2

Gold Standard Comparison

Output is compared against professionally tagged, standards-body-published reference PDFs. We measure how closely our automated structure matches the gold standard — heading hierarchy, table structure, reading order.

3

Diverse Document Regression Suite

Pipeline changes are tested against documents that break most OCR tools — Arabic RTL, Japanese vertical text, degraded mimeographs, multi-column newspapers, and mixed-language pages.

Status: These validation layers are in place and running. We're still improving the pipeline to pass all checks across every edge case — the goal is full compliance on every document, every time.

What's Next

This concept is actively going through testing now.
We're refining the pipeline, working through edge cases,
and improving output quality with every iteration.

We're looking for Ares sites to test with.

If your institution is working toward Title II compliance
for course reserves, we'd love to hear from you.

Atlas Systems · Experimental Concept