# Tesseract OCR vs. AWS Textract (GovCloud)

## Overview

Tesseract OCR is a free, open-source software tool used to convert images with printed text into digital text. It works well for simple documents with clear, typed text, but it struggles with handwritten content and complex layouts.

AWS Textract is a managed service provided by Amazon Web Services that extracts text from scanned documents and images. It can also identify and extract text from handwriting and is designed to scale automatically for large workloads.

AWS Textract (GovCloud) is designed to meet government compliance requirements, offering enhanced security, reliability, and performance for sensitive document processing.

## Cost Comparison

AWS Textract (GovCloud):

- First 1 million pages/month: $1.50 per 1,000 pages

- Over 1 million pages/month: $0.60 per 1,000 pages

- Your usage: 500,000 pages/month = ~$750/month or $3,750 over 5 months

Tesseract OCR:

- Free software

- Hosting costs (e.g., AWS EC2 instance ~ $30/month)

- Additional infrastructure, engineering, maintenance, and scaling costs

## Benefits of Using AWS Textract Over Tesseract

- Higher accuracy, especially with scanned and handwritten documents

- No need to build or maintain server infrastructure

- Scales automatically with document volume

- Designed for government security and compliance (FedRAMP, ITAR)

- Can be integrated easily with other AWS services like Lambda and S3

- Better support and reliability for enterprise-grade use

## Compliance Requirements Explained

FedRAMP (Federal Risk and Authorization Management Program):

- FedRAMP is a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.

- It ensures cloud services used by federal agencies meet strict security standards.

- FedRAMP is required for any cloud service that handles federal data.

More information: https://www.fedramp.gov/

ITAR (International Traffic in Arms Regulations):

- ITAR is a set of U.S. regulations that control the export and import of defense-related articles and services.

- ITAR requires organizations to prevent unauthorized access to defense and military technologies.

- ITAR compliance is crucial for any company dealing with sensitive military or defense data.

More information: https://www.pmddtc.state.gov/

## Use Cases

AWS Textract is ideal for:

- Government agencies processing public records or forms

- Legal and finance teams scanning contracts and invoices

- Healthcare providers digitizing patient records or prescriptions

- Companies with large archives of scanned documents

Tesseract is suitable for:

- Lightweight projects or prototypes

- Documents with simple, clear text only

- Offline environments or disconnected systems

## Final Recommendation

Given the high document volume (500,000 pages/month), the need for strong compliance, and the desire to reduce infrastructure management, AWS Textract is the recommended choice. It offers significantly better accuracy and performance while freeing up your team to focus on more important tasks.

Tesseract may appear cheaper, but the hidden cost of maintenance, lower accuracy, and manual scaling makes it less suitable for large-scale or critical document processing.

## Pros and Cons Comparison

|  |  |  |
| --- | --- | --- |
| Feature | AWS Textract (GovCloud) | Tesseract OCR |
| Accuracy | High (printed + handwritten) | Good for printed only |
| Scalability | Automatic | Manual, needs setup |
| Setup & Maintenance | No maintenance required | High, needs regular support |
| Security & Compliance | GovCloud - FedRAMP, ITAR | None built-in |
| Integration | Works with AWS services | Manual integration needed |
| Cost Transparency | Pay-as-you-go | Free software, hidden ops cost |
| Speed | Faster with high loads | Slower without tuning |
| Structured Extraction | Line & word-level, smart layout | Raw text only |