

Service Name

Amazon Textract

Service Overview

Amazon Textract is a machine learning service that automatically extracts text, handwriting, layout elements, and structured data from scanned documents and images. It goes beyond simple optical character recognition (OCR) to identify, understand, and extract data from forms and tables without requiring templates or configuration. The service uses deep-learning technology developed by Amazon's computer vision scientists and provides APIs for document analysis, expense processing, ID document analysis, lending document processing, and custom queries. Amazon Textract returns extracted data with bounding box coordinates and confidence scores, making it available to developers without requiring machine learning expertise.

Key Use Cases

- **Creating intelligent search indexes:** Using Amazon Textract to create libraries of text detected in image and PDF files for searchable document repositories
- **Automating data capture from forms:** Extracting structured data from forms and building extraction capabilities into existing business workflows for user-submitted data
- **Accelerating data capture and normalization:** Extracting text and tabular data from financial documents, research reports, medical notes, and other varied document sources
- **Intelligent text extraction for NLP:** Providing controlled text grouping as input for natural language processing applications with words, lines, and table cell organization
- **Automating document classification and processing:** Using Analyze Lending API to automatically classify lending documents and route them to appropriate analysis operations

Features

- **Optical Character Recognition (OCR):** Automatically detects printed and handwritten text from documents and images, handling various fonts, styles, and noisy or distorted text
- **Form Extraction:** Detects key-value pairs in document images automatically while retaining context without manual intervention
- **Table Extraction:** Preserves composition of data stored in tables during extraction, maintaining relationships between rows and columns

- **Query-based Extraction:** Allows specification of data to extract using natural language questions without knowing document structure
- **Custom Queries:** Enables customization of pretrained Queries feature using business-specific documents with as few as ten sample documents
- **Signature Detection:** Provides ability to detect signatures on any document or image with location and confidence scores
- **Layout Analysis:** Extracts layout elements such as paragraphs, titles, lists, headers, footers from documents
- **Analyze Expense:** Extracts vendor names, invoice numbers, item prices, total amounts from invoices and receipts with normalized key names
- **Analyze ID:** Extracts information from US government identity documents like passports and driver's licenses without templates
- **Analyze Lending:** Managed API that automates extraction from mortgage loan documents with prebuilt ML models for classification and splitting

Value Proposition

Amazon Textract offers compelling advantages over traditional OCR solutions and manual data entry processes. It eliminates the complexity of building text detection capabilities by providing powerful and accurate analysis through simple APIs, requiring no computer vision or deep learning expertise:

- The service provides scalable document analysis that can process millions of documents quickly to accelerate decision making
- Unlike basic OCR solutions that extract simple text, Textract maintains relationships between data elements like key-value pairs and table structures
- It offers low-cost, pay-per-use pricing with no minimum fees or upfront commitments, plus a free tier for getting started
- The service is always learning from new data with continual feature additions, and supports both synchronous processing for low-latency applications and asynchronous operations for multi-page documents

Service Integration

Amazon Textract integrates seamlessly with core AWS services to enable comprehensive document processing workflows. It integrates with Amazon S3 for document storage and retrieval, supporting both input and output operations:

- Amazon SNS and Amazon SQS are used for asynchronous processing notifications and job completion management
- AWS Lambda enables serverless processing workflows and automated document processing triggers
- Amazon Cognito provides authentication for web and mobile applications using Textract

- The service works with AWS CloudTrail for API call logging and monitoring
- Amazon EventBridge can trigger document processing workflows
- AWS Step Functions orchestrate complex document processing pipelines
- Amazon DynamoDB and Amazon RDS store extracted data and metadata
- AWS IAM manages access control and permissions
- The service also supports VPC endpoints through AWS PrivateLink for private connectivity without internet gateways

Onboarding and Offboarding

Getting started with Amazon Textract requires an AWS account and basic setup steps. Customers can begin by creating an AWS account and configuring IAM users with appropriate Textract permissions:

- The AWS CLI and SDKs need to be installed and configured with access credentials
- Users can immediately try the service through the Amazon Textract Management Console with their own documents, or explore the demonstration features
- The service offers comprehensive documentation including a developer guide, API reference, and step-by-step getting started guide
- Code samples and tutorials are available on GitHub to help developers integrate Textract into applications
- The service provides SDKs for multiple programming languages including JavaScript, Python, Java, and others
- A free tier allows processing up to 1,000 pages monthly with certain APIs, enabling risk-free evaluation and learning

Getting Started Guide: <https://docs.aws.amazon.com/textract/latest/dg/getting-started.html>

Data Removal

Amazon Textract provides clear data removal processes for customer offboarding. Document and image inputs are not stored by Amazon Textract unless explicitly requested by customers. When data is stored, customers can delete images and documents using the AWS Management Console or AWS CLI. For Custom Queries adapters, customers can delete both individual adapter versions and entire adapters using the DeleteAdapter and DeleteAdapterVersion API operations, which remove all associated data and ARNs. Customer content used for adapter generation is deleted once training completes and is not retained for debugging purposes. Customers retain full ownership of their content throughout the process.

Backup/Restore and Disaster Recovery

Amazon Textract operates as a managed service where AWS handles infrastructure backup and disaster recovery. The service itself does not require customer-managed backups as it is stateless for processing operations. However, customers should implement backup strategies for their input documents stored in Amazon S3 and any extracted data stored in databases. AWS Backup can be used to protect associated resources like S3 buckets containing documents and RDS databases storing extracted information. The service provides high availability through AWS's global infrastructure with automatic failover capabilities built into the managed service architecture.

Backup Capabilities

Amazon Textract does not have traditional backup capabilities as it is a stateless processing service. The service processes documents and returns results without storing customer data long-term unless explicitly configured:

- For Custom Queries adapters, the trained models are maintained by AWS but can be deleted by customers when no longer needed
- Customers should implement backup strategies for their source documents in S3 and extracted data in downstream storage systems
- AWS Backup can protect related resources but does not directly backup Textract processing results since they are transient

Disaster Recovery

Amazon Textract supports disaster recovery through AWS's global infrastructure and multi-region availability. The service is available in multiple AWS regions allowing customers to process documents in different geographic locations for redundancy:

- As a managed service, AWS handles the underlying infrastructure disaster recovery including automatic failover and data replication
- Customers can implement cross-region strategies by processing documents in multiple regions and storing results in geographically distributed storage systems
- AWS Elastic Disaster Recovery does not directly apply since Textract is a stateless managed service

Recovery Objectives

Amazon Textract helps customers achieve recovery objectives through its managed service architecture and multi-region availability. The service provides near real-time processing for single-page documents, supporting low RTO requirements. For batch processing scenarios, asynchronous operations can be restarted quickly in alternate regions if needed. Since the service is stateless, RPO objectives depend on backing up source documents and extracted results rather than the processing service itself.

Customers can achieve sub-hour RTO by maintaining document repositories in multiple regions and using cross-region replication for critical processing workflows.

Service Constraints

Service Quotas: <https://docs.aws.amazon.com/general/latest/gr/textract.html>

FAQ: <https://aws.amazon.com/textract/faqs/>

Technical Requirements

Service Documentation: <https://docs.aws.amazon.com/textract/latest/dg/what-is.html>

User Guide: <https://docs.aws.amazon.com/textract/latest/dg/what-is.html>

API Reference: https://docs.aws.amazon.com/textract/latest/dg/API_Reference.html

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: <https://aws.amazon.com/textract/pricing/>

Cost Optimization

Amazon Textract offers several unique cost optimization opportunities through its pay-per-page pricing model. Customers can optimize costs by preprocessing documents to remove unnecessary pages, using appropriate API features (basic OCR vs:

- advanced analysis), and implementing confidence score thresholds to reduce manual review costs
- Asynchronous processing enables batch operations that can be more cost-effective than real-time processing for large volumes
- The service offers volume-based pricing tiers with reduced costs after one million pages for most APIs
- Custom Queries can improve extraction accuracy, reducing downstream processing costs and manual review requirements
- Customers should evaluate which combination of Features (Forms, Tables, Queries) they actually need rather than using all features by default

Savings Considerations

Main cost savings with Amazon Textract come from eliminating manual data entry and document processing labor costs. The service can process thousands of documents per hour compared to manual processing, providing significant operational savings:

- Volume-based pricing provides cost reductions for high-usage scenarios, with prices decreasing after processing one million pages monthly
- The free tier offers 1,000 pages monthly for text detection, 100 pages for document analysis, and other allowances for evaluation without cost
- Preprocessing optimization like document quality improvement can increase extraction accuracy, reducing costly manual review processes
- Using appropriate APIs for specific use cases (Expense API for invoices vs general document analysis) provides better cost efficiency and accuracy than generic approaches