

A richer source of regulatory truth.

Waymark Tech enables rapid impact assessments and compliance workflows with Adobe PDF Extract API.



Established

2016

Employees: 9
London, United Kingdom
www.waymark.tech

20x

Improvement in speed of
extractions compared to
competitive solutions

Products:

[Adobe Document Services >](#)
[Adobe PDF Extract API >](#)

Objectives

Cut complexity of extracting information from documents with key data in both text and tables

Accelerate time-to-notification to give clients the latest news and regulatory changes

Reduce time spent by Waymark team handling common errors due to insufficient text or table recognition

Results

Delivers simplicity of one solution to eliminate 5+ point products for data and content extraction tasks

Processing thousands of PDF files per day instead of just 50 to 100

Increased accuracy with manual exception handling rate dropping 80% to 90%

New regulations are the lifeblood of Waymark Tech's business. The company's global clients, from government policymakers and regulators to compliance supervisors, professional services firms, and financial institutions, all depend on Waymark to deliver fast and accurate information about regulatory changes.

Companies in heavily regulated industries, such as financial services, face a particularly daunting challenge of constant regulatory changes in each market where they operate. Complexity multiplies quickly, considering new regulations or updates can be issued by a different state, province, country, or supra-national regulators.

Regulatory announcements can arrive with little time for companies to react. Each new announcement sets off a flurry of activities for affected businesses, which need to identify, assess, and potentially address associated risks, obligations, and other impacts. "Compliance officials need to be able to quickly confirm whether the particulars of a regulatory change may affect their firm, enabling them to start mitigating any reputational and operational liabilities as soon as possible," says Mark Holmes, CEO, Waymark Tech.

An innovative startup on an impressive growth trajectory, Waymark simplifies regulatory information sharing and compliance operations for its clients with an array of tools on its Wayfinder platform. The platform uses artificial intelligence and machine learning, including natural language processing, to power capabilities for regulatory notifications, searching, and analytics, plus workflow and collaboration tools for regulatory change management.

[Adobe PDF Extract API](#), a web service powered by [Adobe Sensei](#) machine learning to unlock the structure and content elements of any PDF, serves as Waymark's primary engine for extracting information from regulatory announcements and related documentation in PDF. Waymark had tried numerous point products for different segments of the information extraction process, but these created many time-consuming, imperfect steps.

"There are open source options that give you bits of the needed information. One of them might be really good with metadata information about the text, for example, but then it will fail when it encounters a table or an image," says Sean Gilley, Lead Engineer, Waymark Tech.

"Adobe PDF Extract API identifies tables — even implicit tables — very readily, which is quite impressive. It's a game changer because we can also use it to take a table from a PDF and export it to CSV format. That's extraordinarily useful."

Sean Gilley
Lead Engineer, Waymark Tech

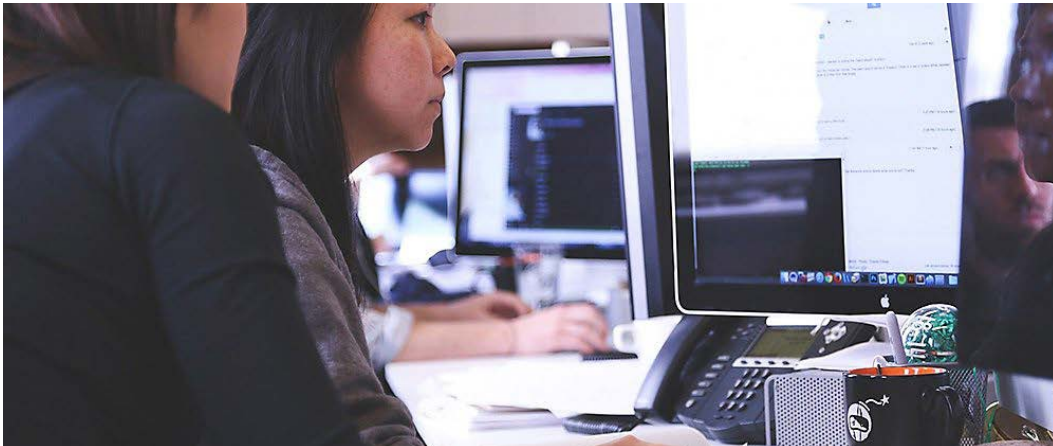
A bountiful harvest of regulatory data

Waymark selected Adobe PDF Extract API because it provides a valuable, more holistic approach to automating data harvesting. The service is part of the [Adobe Document Services](#) toolkit of cloud-based REST APIs and SDKs, providing flexible integration and productivity possibilities so developers can invent new automated document workflows.

A comprehensive PDF extraction service, PDF Extract API provides accurate extraction of all the elements in a source document. "It's good to have a range of extraction tools for our application sitting in a single API with Adobe PDF Extract API, rather than having to juggle more than five tools for certain tasks, like we did in the past," says Gilley.

He finds the service especially valuable for extracting complex tables that often appear in regulatory documents. "Adobe PDF Extract API identifies tables — even implicit tables — very readily, which is quite impressive. It's a game changer because we can also use it to take a table from a PDF and export it to CSV format. That's extraordinarily useful," Gilley says.

When Waymark onboards data to its platform, which uses a Python stack with React at the front end, it's not simply performing a bulk upload of all the text from a PDF file. First, the company takes JSONs and other file outputs from PDF Extract API using the Python SDK and imports them to its own proprietary format. "The Python API is really simple to use," says Gilley.



Refined intelligence better illuminates regulations

The Waymark object then stores the various content elements and structural properties of a document down to the paragraph or sub-paragraph level, recording the relationships between these items. To accomplish this, Gilley's team uses the bounding box information from Adobe PDF Extract API to make assumptions regarding document content. The team sets up pipelines that account for the position of a paragraph within a chapter or section of a document, for example.

"The metadata around each of the elements within a bounding box is vital because then you can start building parsers for different sources and set up rules, such as a subtitle for a document from Source A will always be in this font size, this typeface, indented this much, and underlined. The sophisticated intelligence that Adobe PDF Extract API delivers is very useful, providing a much easier way to harvest truly meaningful text, metadata, and related content from a PDF," says Gilley.

The superior capabilities of Adobe PDF Extract API for recognizing tables and key contextual data makes it possible for Waymark to automate the processing of thousands of PDFs a day. In its earlier days, during spikes in document intake volume, the company hired temporary staff to manually deal with the high volume of exceptions coming from unrecognized elements. This proved to be an expensive approach, prone to data entry errors, and resulting in only 50 to 100 completed PDFs per day.

"The metadata around each of the elements within a bounding box is vital because then you can start building parsers for different sources and set up rules, such as a subtitle for a document from Source A will always be in this font size, this typeface, indented this much, and underlined. The sophisticated intelligence that Adobe PDF Extract API delivers is very useful, providing a much easier way to harvest truly meaningful text, metadata, and related content from a PDF."

Sean Gilley
Lead Engineer, Waymark Tech

Higher volume capacity helps Waymark keep up with one of its principal data sources, EUR-Lex. The official website for European Union law and related public documents, it publishes regulatory documents in all of the European Union's 24 official languages.

"Now we can take thousands of PDFs and just drop them into an AWS S3 bucket, queue them up with Adobe PDF Extract API, then direct all the output files to another S3 bucket for further processing later. It's much faster, more precise, and efficient. These days it's just a matter of occasional exception handling for my team. It's no longer the norm," Gilley says.

Gilley estimates exception rates have fallen by 80% to 90% versus competitive extraction solutions Waymark tried in the past, due to more refined recognition and intelligence features in Adobe PDF Extract API. In addition, he estimates Waymark has seen an improvement in data extraction speeds by at least 20 times with the Adobe service compared to alternative solutions the company previously used.

Accelerating client communications and internal innovation

Waymark clients appreciate the company's speed. "We have customers who say, 'If I don't know about a regulation change before a law firm partner has told my boss about it, then it looks bad for me.' We are scanning our principal regulatory information sources every 15 minutes. We can alert clients immediately if a new item meeting their search criteria shows up on our platform," says Holmes.

The efficiency gains promote innovation within Waymark. "Adobe PDF Extract API frees up our IT team's time for many other important priorities, like developing enhancements to our Wayfinder platform that promote the growth of our company," says Atta Rehman, CTO, Waymark Tech.

The Adobe solution helps unlock new business development opportunities for Waymark, enabling the fast-growing, nimble company to deliver greater value for clients by enhancing its Wayfinder platform with new services.

In the future, Waymark envisions allowing its clients to use its Waymark platform on a self-service basis, securely onboarding their own policy documents related to regulatory compliance, with PDF Extract API helping underpin the service. With this, clients could apply Waymark's data enrichment services to their own policy documents, including natural language processing and analytics that apply intelligence across both their internal policy documents as well as the relevant regulations and third party publications in Waymark's database, to deliver more complete assessments and streamlined workflows.