Adobe

ADOBE PDF EXTRACT API

Extracting exam questions, automatically.

Cambridge Assessment intends to digitize more than 100 years of test material with Adobe PDF Extract API.

Cambridge Assessment

Established
1858
Employees: 3,000
Cambridge, UK
www.cambridgeassessment.org.uk

90%
Accuracy rate while maintaining complex formatting

Products:
Adobe Document Services ›
Adobe PDF Extract API ›

Objectives

Support growing digital education needs

Tap into value of 160 years of test content for educators worldwide

Automate content harvesting and banking processes while maintaining accuracy

Results

Preserves visual elements and structure of exam questions

90% accuracy rate while maintaining complex formatting

Processes a 40-question exam in as little as **40 seconds**

Saves 2,000 days of labor for every 50,000 questions harvested and eliminates costs for manual data entry

Every year, more than 8 million learners in over 170 countries demonstrate what they have learned through tests and examinations designed and delivered by Cambridge Assessment. The organization offers a wide range of qualifications, from AS/A Levels, GCSEs, and other exams aimed at students aged 5 to 19, to English proficiency certifications recognized by universities, employers, and government bodies in 130 countries.

Education systems have changed dramatically since Cambridge Assessment held its first examinations in 1858. But many students still take its exams the way that they did 160 years ago: with pencil and paper.

"The results from Cambridge Assessment examinations can help determine university placements or job qualifications," says Terry Child, Principal Product Manager at Cambridge Assessment. "Security is top priority. In many ways, paper testing is still one of the most secure methods available."

However, Cambridge Assessment recognizes that this traditional, paper-based testing model is slowly changing as education grows more digital. More schools and educators are delivering lessons, homework, and midterms digitally.

Cambridge Assessment is making huge strides in digital education by leveraging its wealth of exam history and knowledge. By harvesting questions from past exams, the company is creating a content bank filled with unmatched levels of educational content. Cambridge Assessment has big plans for its digital database, from fueling classroom content to using artificial intelligence (AI) to analyze and understand test performance. The more historical exam data that Cambridge Assessment adds to its digital database, the more valuable its service becomes for customers around the world.

While many exams are archived in PDF, the process of extracting individual questions from an exam and entering them into the content bank was previously very manual. "While there are many tools on the market that will turn PDF data into a string of text, we had difficulty finding a method that would identify the formatting, figures, tables, and question structure that we needed to create our database," says Child. "Adobe PDF Extract API changes things. It's the only tool we've found that will allow us to maintain question structure and start automating question harvesting."

"Adobe PDF Extract API allows us to extract the context of our questions. This is the first step towards automating how we harvest questions from exams for our content bank."

Terry Child
Principal Product Manager, Cambridge Assessment

Accurately extracting PDF formatting and structure

The Cambridge Assessment team uses the Java SDK to parse test papers in PDF using **Adobe PDF Extract API**. The API turns test papers into JSON output, which Cambridge Assessment post-processes using custom Java code and then transforms into QTI XML format — the standard specification for electronic assessment systems. Once stored in this standard format, the questions and answers can be reused in any compatible system.

While Child looked at other text extraction tools and experimented with various Python libraries, only Adobe PDF Extract API preserves visual formatting and structure along with the text. It distinguishes between text, images, figures, and tables. It identifies bold, italics, subscripts, superscripts, and symbols that frequently appear in advanced math questions.

"If we can't distinguish something like a squared sign in a mathematical formula, then many of our science and math questions are useless," says Child. "Even simple formatting like font weight is critical. Our questions have been through countless review processes. If one word in a question is bolded, that's because a team agreed that it was essential to the understanding of the question. Capturing all of that formatting is part of the question itself."

Cambridge Assessment has seen impressive accuracy and speed from its tests of **Adobe PDF Extract API**. "We're seeing hit rates above 90%, after the Cambridge Java code has post-processed, on multiple choice questions with Adobe PDF Extract API," says Child. "We can extract all questions from a 40-question examination in just 40 to 90 seconds."

While the ability to properly extract formatting from a PDF file is critical, the ability to accurately parse tests into individual questions is equally important. Adobe PDF Extract API provides the information that Cambridge Assessment needs to identify where each question starts and ends. Child used this information to create an automated process that separates a single test paper into a series of questions.

"Adobe PDF Extract API allows us to extract the context of our questions," says Child. "This is the first step towards automating how we harvest questions from exams for our content bank."

"By using Adobe PDF Extract API to automate how we pull questions into the content bank, we will save over 2,000 days of labor for every 50,000 questions harvested and eliminate the costs of hiring temporary workers for data entry."

Terry Child
Principal Product Manager, Cambridge Assessment

Harvesting multiple-choice questions with speed

Cambridge Assessment started testing **Adobe PDF Extract API** on the most standardized type of exam question: multiple choice. Each multiple-choice question starts with the context of the question, which may include several statements, figures, or tables. This is followed by the question prompt and four possible responses.

Adobe PDF Extract API pulls all PDF information into JSON format while also providing associated png and csv data for images and table respectively. Cambridge Assessment created a post-processing pipeline that applies a set of logic rules to the JSON output to separate each question into the context, prompt, and responses.

Once delivered into QTI XML format, a subject expert checks the question for errors and adds metadata to help categorize each question according to what it's testing, difficulty level, and other such information. Finally, the questions stored in a content bank for use by customer-facing products and services.

Before working with Adobe PDF Extract API, entering questions into the content bank was a tedious and manual process. Cambridge Assessment hired temporary workers to either retype questions or copy and paste content from PDF files into QTI XML format files. This manual process was prone to errors and slow.

"By using Adobe PDF Extract API to automate how we pull questions into the content bank, we will save over 2,000 days of labor for every 50,000 questions harvested and eliminate the costs of hiring temporary workers for data entry," says Child. "With the time saved, we will be able to harvest more questions and build a much richer content bank."

"We have a huge bank of intellectual property in our test archives that we can harness. If we can digitize it all, we can create an unmatched resource for educators and learners."

Terry Child
Principal Product Manager, Cambridge Assessment

Leveraging 160 years of testing resources

Cambridge Assessment will start by harvesting multiple-choice questions, but Child plans to create rules sets to import other types of questions — such as short-answer and essay questions — as soon as possible. Child also hopes to potentially add more content from the organization's 160 years of testing history by scanning historic exam papers, converting them to QTI, and adding the questions to the content bank.

Currently, employees must add metadata to each question by hand. But that, too, might be automated in the future. Once the content bank has more questions added, Cambridge Assessment wants to start training AI to understand questions and add appropriate metadata. If successful, AI will add another layer of automation for even faster question harvesting.

The enriched content bank will be ripe for a wide variety of content-as-a-service offerings in the future. Cambridge University Press might take information from the content bank for use in textbooks or test preparation materials. It could even create a self-service system where educators could instantly create their own tests for any learning system anywhere in the world. This could help educators more easily achieve personalized learning options tailored to the needs of specific schools, classrooms, or even individual learners.

"We have a huge bank of intellectual property in our test archives that we can harness," says Child. "If we can digitize it all, we can create an unmatched resource for educators and learners. **Adobe PDF Extract API** gives us the functionality that we need to automate processes and provide access to a greater variety of test questions quickly."