CASE STUDY

iText pdf2Data: enabling intelligent, efficient extraction of PDF payslip data

iText pdf2Data: allowing PayDashboard's interactive digital payslip platform to process PDF output from legacy payroll software to extract the necessary data and present it in their responsive web application.

BACKGROUND

PayDashboard is a UK-based company offering digital payslip solutions. They have developed a platform that allows employees to access their pay data through a user-friendly, secure online portal. PayDashboard's solution integrates with existing payroll software to deliver their SmartSlip online payslips and other payroll data such as PAYE forms, P60, P45, and P11d documents, in a more modern way.

Fully accessible on mobile devices, PayDashboard provides intelligent dashboards to display workforce and pay analytics and graphs in an attractive and easy to read way. By offering modular functionality, their clients can pick and choose the features they need to make PayDashboard work better for their business, whether they require a standalone payslip portal or want to integrate digital payslips into their existing HR platform.

By repurposing and reimagining simple payroll data into a slick, responsive and modern interface, its portal enables much more than simply numbers on a screen or a piece of paper. Though digital payslips of some form are now common in the UK and other countries, you often get little more than a simple representation of your pay data.

Even when you have a PDF payslip, its rarely the case that whatever system created the payslip took advantage of the advanced capabilities offered by the PDF format. Frequently, the best-case scenario for such payslips is that the data is machine-readable, while at worst it is just a scanned image embedded in a PDF container. By contrast, PayDashboard's payslip portal offers multiple ways to visualize pay data, allowing employees to better understand their pay and their payslip, while giving greater access to payroll insights with interactive dashboards

Since PayDashboard's solution first went live around five years ago they have gained a reputation for excellence in the payroll industry. PayDashboard's customers range from small to medium enterprises (SMEs) to large private and public sector employers, accountants, and payroll bureaus. Being ISO 27001 certified and GDPR compliant, with all data encrypted in accordance with industry best practices means PayDashboard has become a leader in their field, and in 2019 joined the UK Government's G Cloud 11 framework which enables Public Sector organizations to easily utilize PayDashboard's SmartSlip technology through its Digital Marketplace.

An example representation of PayDashboard's user-frtiendly payslip portal

GOALS

For some time, PayDashboard had been looking at ways to intelligently extract data from PDF in a structured and easily digestible form. When talking to prospective clients, they were often using payroll systems which for various reasons could not connect to PayDashboard's API. As Luke Hopton, the Engineering Manager at PayDashboard notes "We'd already built a number of different CSV and XML formats into our API to understand and ingest the data from payroll systems. However, sometimes this simply wasn't possible and we kept finding that the common denominator of all these systems was that they produced PDF payslips. So, being able to get the data out of those PDF payslips and into our platform to do all the interesting things we can do with it was essential".

Over the last few years, electronic payslips have become commonplace, and many employees receive their payslips as a PDF. Many of PayDashboard's larger customers are accounting firms or payroll outsourcers who run payrolls for potentially thousands of separate employers. As such, they may be using advanced payroll software which can connect directly with the PayDashboard platform's API. However, many smaller businesses do not have the technology or resources available to achieve this, and even larger companies and payroll providers may only be able to generate PDF payslips. As Luke explains, "Big payroll companies weren't interested in writing one-off custom CSV or XML export functionality for us, so until we started using iText pdf2Data we had no way to solve the problem of getting the data out of PDFs."

In order to allow such companies and their employees to benefit from the value proposition PayDashboard offers, the development team at PayDashboard began looking into solutions that would enable them to extract the required data directly from a PDF payslip.

After looking at a number of options, they came across iText pdf2Data which allows the intelligent recognition and extraction of data from PDFs using a predefined template, which can be configured using its intuitive browser-based template editor. Since PDF documents like payslips have a consistent structure and appearance, with the only changes from document to document being the actual data they contain – the employee's name, National Insurance number, and of course the pay data itself – PayDashboard's team could use an example payslip to define a template for each of their clients, and then automatically process each subsequent payslip using the iText pdf2Data SDK.

CHALLENGES

Such a solution had to be performant, and able to handle large amounts of data, as the PDF they receive from a client may contain the pay data for many thousands of employees.

SOLUTION

What initially attracted PayDashboard to iText pdf2Data was Luke's prior experience with the iText PDF library. "I was aware of iText as I had used it to generate PDF reports from code in a previous role. When we found iText pdf2Data, knowing it had your backing gave it a lot more credibility since I knew you were a pretty big player in PDF".

One thing that really stood out to PayDashboard's team was iText pdf2Data's template editor. During their research they had found various tools to scrape data out of PDFs. However, being code-based, these tools tended to simply output data as an unstructured JSON dump for example. In contrast, being able to simply draw the required selectors onto a PDF was a big selling point. As Luke puts it "Having the template editor was I think the big selling point for us to take to the proof-of-concept stage, as we could actually just draw the selectors onto a PDF and say" give me that, that's the bit I'm looking for".

"I was aware of iText as I had used it to generate PDF reports from code in a previous role. When we found iText pdf2Data, knowing it had your backing gave it a lot more credibility since I knew you were a pretty big player in PDF."

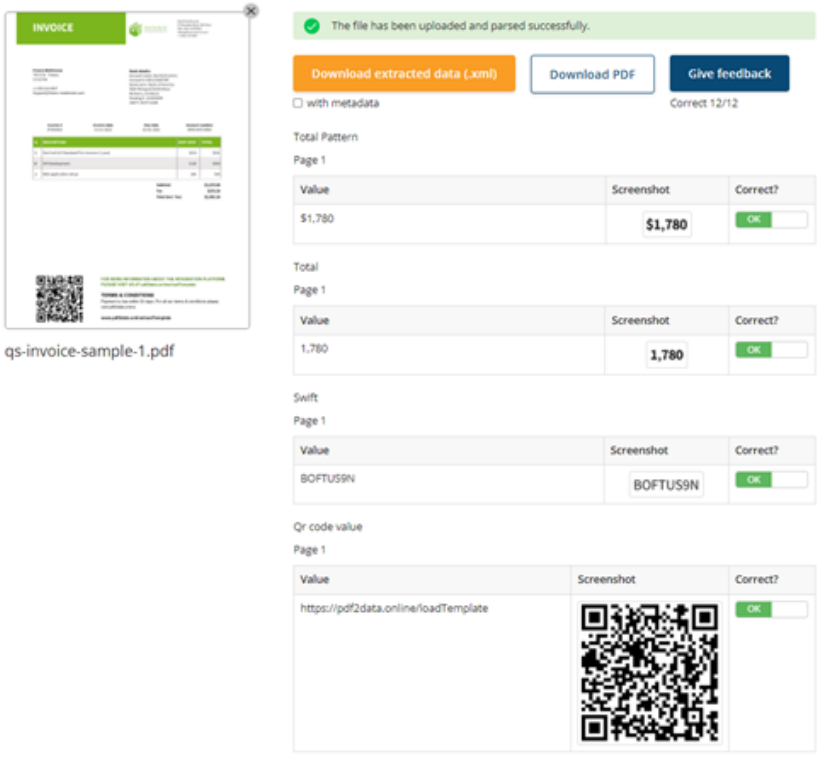
LUKE HOPTON, ENGINEERING MANAGER

HOW IT WORKS

iText pdf2Data's template editor lets you define areas and rules in a template which correspond to the content you want to extract. The template can then be visually validated with other documents to confirm data is recognized correctly, before being parsed by the pdf2Data SDK to process all subsequent documents matching that template.

iText pdf2Data offers a wide range of different selectors which can recognize things like dates, fonts, paragraphs, tables and much more, meaning it can be used for all kinds of PDF documents which use consistent structure and formatting, like payslips, invoices, forms etc. When you want to create a template from a PDF, you simply draw a box around the data you want to extract, and then use one of the available selectors to define what the data is so it can be extracted correctly.

Once the rules for a template have been defined, the data from the subsequent documents is extracted in XML format which can be quickly and easily processed and repurposed as required.



The iText pdf2Data's template editor showing data extracted from an invoice template

"Having the template editor was I think the big selling point for us to take to the proof-of-concept stage, as we could actually just draw the selectors onto a PDF and say" give me that, that's the bit I'm looking for".

LUKE HOPTON, ENGINEERING MANAGER

RESULT

Going to using iText pdf2Data in a production environment was achieved remarkably quickly, with the team implementing the initial proof-of-concept within a single three-week development sprint. Luke says, “We built up an example mapping of data from a payslip generated from one of the main payroll systems many of our clients use, just to prove we could work out the basics. From that, we simply did some fine-tuning to handle more complex parts of the document and make it production-ready”.

Since beginning to work with iText pdf2Data in 2019, PayDashboard’s team are very happy with its PDF data extraction prowess, the support from iText with their implementation, and its regular updates with new or improved functionality.

As a result of their positive experience with iText pdf2Data, PayDashboard are also looking at implementing other iText PDF functionality into their platform. As Luke explains. “We are looking at modernizing the way we generate downloadable PDF payslips from our platform. Sometimes people need a PDF version to prove their income, for example to a bank. We create a simple web page which is converted to PDF using an open-source plugin. However, this plugin is no longer maintained so we think iText would be a great way to replace it.”

ABOUT PAYDASHBOARD

PayDashboard is a truly independent interactive payslip platform for employees, employers, professional advisors, and partners. By integrating with existing payroll software, it allows their clients to provide employees with digital payslips via a secure online portal. Providing pay data in a digital format unlocks a wealth of innovation that is just not possible with a PDF payslip.

Through their portal they deliver additional services that add value for their clients. From financial education to help employees to understand their pay and finances, to employer dashboards to track and analyze workforce costs, PayDashboard is all about adding value at what is traditionally the end of the payroll process.

If you’re interested in learning how **iText pdf2Data** could help you with your own PDF data extraction challenges, we have an online demo showing its template editor and data extraction in action. We have example templates to play around with, or you can even upload your own PDF!

ABOUT US

iText is a global leader in innovative award-winning PDF software. It is used by millions of users - both open source and commercial - around the world to create digital documents for a variety of purposes: invoices, credit card statements, mobile boarding passes, legal archiving and more. iText works and works well. Our customers choose iText because of our world-class quality of software, and our reliable mature, proven technology in the iText code library iText 7 Suite. We are recognized as a global thought leader and innovator in PDF solutions and functionalities. As an open source code library, iText PDF can be embedded into the document solution workflows of various industries and their applications.

Europe, Middle East, Africa & CIS

AA Tower
Technologiepark - Zwijnaarde 122
9052 Zwijnaarde
Belgium

sales.isb@itextpdf.com
www.itextpdf.com
+32 9 298 02 31

Americas

530 Harrison Ave,
Second Floor
Boston, MA 02118
United States

sales.isc@itextpdf.com
www.itextpdf.com
+1 617 982 2646

Asia & Oceania

Republic Plaza
9 Raffles Place, Level 6,
Republic Plaza 1
Singapore 048619

sales.isa@itextpdf.com
www.itextpdf.com
+65 6932 5062