CONVERSION OF PDF DOCUMENTS TO EXCEL

Mr. Y Kumara Siva Datta (Author) Final year M-tech pursuing in Computer Science & Engineering,Department of PG Studies, Visveswarya Technological University, Mysuru.

Mrs. Shashirekha .H (Co-Author) Assistant Professor, Department of Computer Science & Engineering, Department of PG Studies, Visveswarya Technological University, Mysuru

Abstract: One of the primary challenges facing in the online Pdf to Excel convertor is managing and processing the huge documents across internal and external server and the security related concerns because of the third-party portals, still in many organizations still often rely on humans to perform manual work to process the information.

Finally, there is a solution that automates Virtual and manual, repetitive, task that needs to Content to ensure it is accessed and acted on as part of critical convertor processes. A new tool made up in helping the convertor task easier without any complex coding.

Gone are the days when Organizations were reliant on physical books of records or with the manual entry where conversion will not gone be easy when we want to convert bulk of files without any error.

Index Terms – PDF Documents, Convertor

I. INTRODUCTION

For accurate data entry from PDF file into various sources like text, word, excel, image, html, handwritten documents, online database, web data, etc. Professional data entry operators are hired at Om Data Entry India for most reliable and top class quality results. We promise to deliver supreme accuracy and fast turnaround time.

There are various features of our PDF to excel data entry services provided to your company, outsource PDF to excel data entry services to Om Data Entry India for converting your accounts, payrolls, balance sheets, customer records or any other data in Excel file formats. Save on budget, resources and time by outsourcing PDF to excel data entry services at our company based in India.

II. EXISTING SYSTEM:

In the Existing System where community had an increasing client base and struggled to maintain a supportive operational infrastructure. Of course, it was essential to keep up the same standard of conversion. The Organization faced numerous labor-intensive challenges many of which stemmed from inefficient on the manual labor of employees. In the current existing online convertor, which just capture the image and it will store it into and excel sheet but we cannot edit those things since those are images and they are not efficient .Many Organization providers suffer financially due to such inefficient or misspelled words.

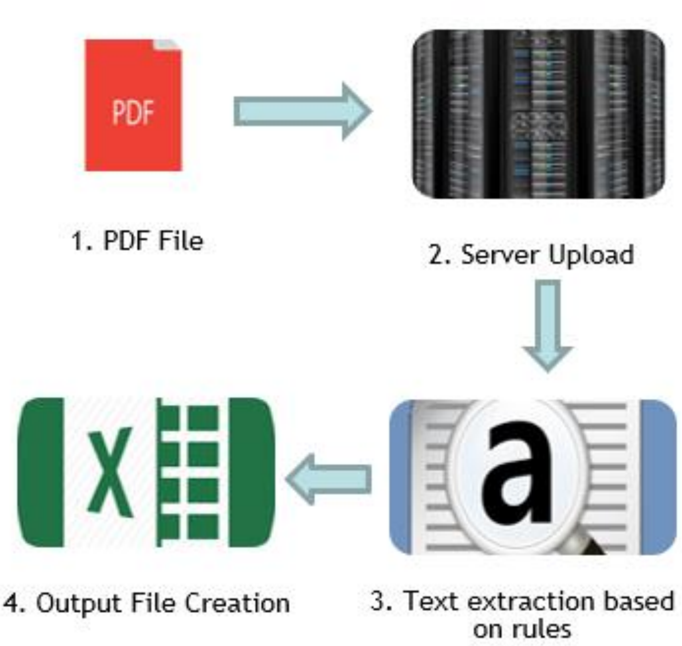
III. PROPOSED SYSTEM

- The PDF file is browsed by the user
- It is uploaded into the server.
- The text is extracted based on the rules.
- The data is then inserted into DB and the output file is created.

IV. OBJECTIVE

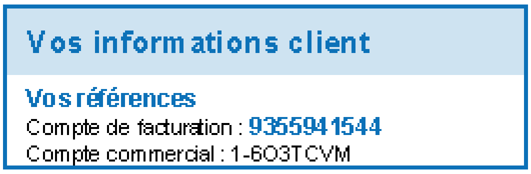
This tool can provide a tremendous opportunity for many organizations where they can clearly pre-define the rules based on the rule the text is been extracted from the PDF and the process is efficient .Cost –savings and improved experience –Likely leading to better outcomes for all

V. POC Solution



- The PDF file is browsed by the user
- It is uploaded into the server.
- The text is extracted based on the rules.
- The data is then inserted into DB and the output file is created
- IDE used is NET Beans 7.4
- The server used to host is Glass Fish 4.0
- JAVA EE 7, JDK 1.7.0 are the JAVA packages
- MYSQL 5.7 is the database used

2 SAMPLE PDF

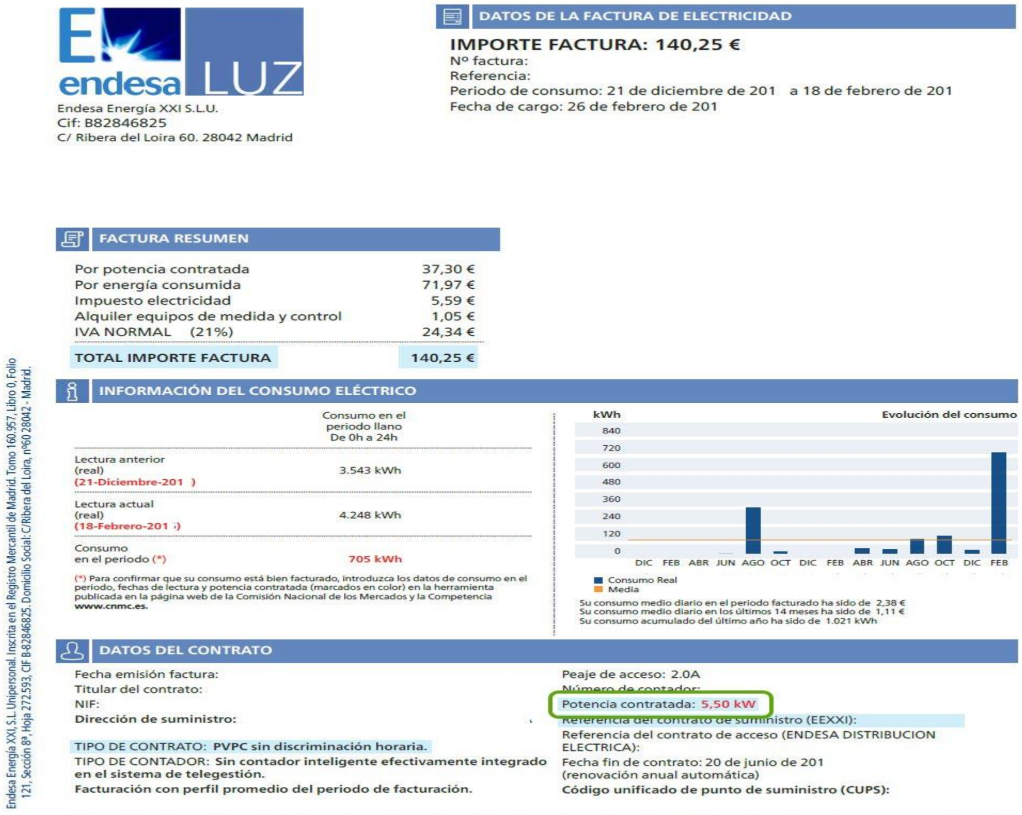


Vos informations client

Vos références

Compte de facturation : 9355941544

Compte commercial : 1-6O3TCVM



Endesa Energía XXI S.L.U.
Cif: B82846825
C/ Ribera del Loira 60. 28042 Madrid

DATOS DE LA FACTURA DE ELECTRICIDAD

IMPORTE FACTURA: 140,25 €
Nº factura:
Referencia:
Periodo de consumo: 21 de diciembre de 201 a 18 de febrero de 201
Fecha de cargo: 26 de febrero de 201

FACTURA RESUMEN

Por potencia contratada	37,30 €
Por energía consumida	71,97 €
Impuesto electricidad	5,59 €
Alquiler equipos de medida y control	1,05 €
IVA NORMAL (21%)	24,34 €
TOTAL IMPORTE FACTURA	140,25 €

INFORMACIÓN DEL CONSUMO ELÉCTRICO

	Consumo en el periodo llano De 0h a 24h
Lectura anterior (real) (21-Diciembre-201)	3.543 kWh
Lectura actual (real) (18-Febrero-201)	4.248 kWh
Consumo en el periodo (*)	705 kWh

(*) Para confirmar que su consumo está bien facturado, introduzca los datos de consumo en el periodo, fechas de lectura y potencia contratada (marcados en color) en la herramienta publicada en la página web de la Comisión Nacional de los Mercados y la Competencia **www.cnmc.es.**

Su consumo medio diario en el periodo facturado ha sido de 2,38 €
Su consumo medio diario en los últimos 14 meses ha sido de 1,11 €
Su consumo acumulado del último año ha sido de 1.021 kWh

DATOS DEL CONTRATO

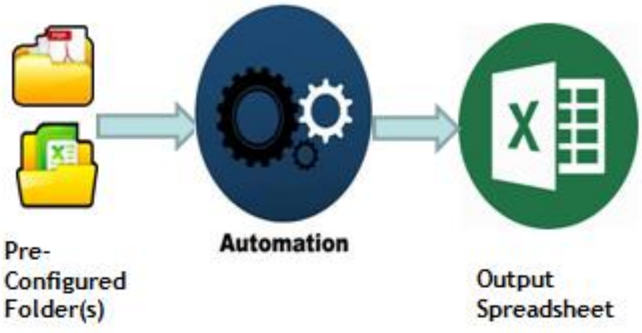
Fecha emisión factura:
Titular del contrato:
NIF:
Dirección de suministro:

TIPO DE CONTRATO: **PVPC sin discriminación horaria.**
TIPO DE CONTADOR: **Sin contador inteligente efectivamente integrado en el sistema de telegestión.**
Facturación con perfil promedio del periodo de facturación.

Peaje de acceso: 2.0A
Número de contador:
Potencia contratada: 5,50 kW
Referencia del contrato de suministro (EEXXI):
Referencia del contrato de acceso (ENDESA DISTRIBUCION ELECTRICA):
Fecha fin de contrato: 20 de junio de 201
(renovación anual automática)
Código unificado de punto de suministro (CUPS):

Endesa Energía XXI, S.L. Unipersonal. Inscrita en el Registro Mercantil de Madrid. Tomo 160.957, Libro 0, Folio 121, Sección 8ª, Hoja 272.593, CIF B-82846825. Domicilio Social: C/Ribera del Loira, nº60 28042 - Madrid.

2.1 Approach



This process is carry into 4 steps

1 .The input folders would be pre-configured and the files placed in them.
2. The application would process the individual files against the defined rules and move the files into a separate file location .
3. The extracted data would be inserted into DB
4. The data would then be fetched from the DB and inserted in to a spreadsheet

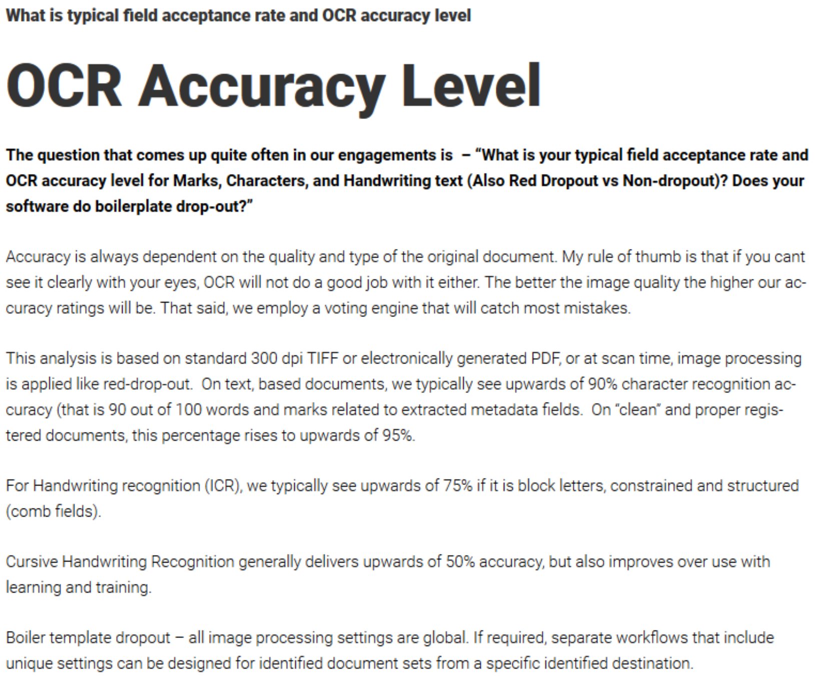
2.2 Advantages

- Accelerated time to value
- Reduced human errors
- Decreased costs

2.3 Scanned/Image based PDF

- Text extraction from scanned/Image based PDF can be performed by OCR tools which are at most 90% accurate
- There are no tools available to verify whether the text extraction is accurate

2.4 TOOL: OCR



What is typical field acceptance rate and OCR accuracy level

OCR Accuracy Level

The question that comes up quite often in our engagements is – "What is your typical field acceptance rate and OCR accuracy level for Marks, Characters, and Handwriting text (Also Red Dropout vs Non-dropout)? Does your software do boilerplate drop-out?"

Accuracy is always dependent on the quality and type of the original document. My rule of thumb is that if you cant see it clearly with your eyes, OCR will not do a good job with it either. The better the image quality the higher our accuracy ratings will be. That said, we employ a voting engine that will catch most mistakes.

This analysis is based on standard 300 dpi TIFF or electronically generated PDF, or at scan time, image processing is applied like red-drop-out. On text, based documents, we typically see upwards of 90% character recognition accuracy (that is 90 out of 100 words and marks related to extracted metadata fields. On "clean" and proper registered documents, this percentage rises to upwards of 95%.

For Handwriting recognition (ICR), we typically see upwards of 75% if it is block letters, constrained and structured (comb fields).

Cursive Handwriting Recognition generally delivers upwards of 50% accuracy, but also improves over use with learning and training.

Boiler template dropout – all image processing settings are global. If required, separate workflows that include unique settings can be designed for identified document sets from a specific identified destination.

CONCLUSION:

The automation of manual task is an important strategy for performance improvements as the industry works to cut costs and improve efficiency
This tool can provide a tremendous opportunity for many organizations where they can clearly pre-define the rules based on the rule the text is been extracted from the pdf and the process is efficient, Cost savings and improved experience -Likely leading to better outcomes for all.

REFERENCES

[1] T. M. Breuel, "Two geometric algorithms for layout analysis," in Document analysis systems v. Springer, 2002, pp. 188–199.
[2] A. C. e Silva, A. M. Jorge, and L. Torgo, "Design of an end-to-end method to extract information from tables," International Journal of Document Analysis and Recognition (IJDAR), vol. 8, no. 2-3, pp. 144–171, 2006.
[3] T. Kieninger and A. Dengel, "A paper-to-html table converting system," in Proceedings of Document Analysis Systems (DAS), vol. 98, 1998.
[4] "Applying the t-recs table recognition system to the business letter domain," in Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on. IEEE, 2001, pp. 518–522.