Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Issue 4/2023

AUTOMATED (SEMANTICS DRIVEN) DATA RETRIEVAL FROM FISCAL DOCUMENTS: A COMPREHENSIVE APPROACH

Vasile MINEA[1], Cornel STAN [2], Gheorghe – Dragoș FLORESCU[3] , Costin LIANU[4], Cosmin LIANU[5]

[1,2,3] Senior Software Agency SRL, Tudor Vladimirescu Ave, No. 45, Floor 1-2, District 5, Bucharest Municipality, Romania, Phone 004021 310 7481;E-mails: vminea@seniorsoftware.ro; cstan@seniorsoftware.ro; dflorescu@seniorsoftware.ro

[4] Spiru Haret University, Faculty of Economic Sciences, 46G Fabricii Str, District 6, Bucharest, Romania, Tel.: +40729868364, Fax: +0213169793, Email: costin.lianu@spiruharet.ro; clianu@gmail.com

[5] Bucharest Academy of Economic Studies, PhD student, Roman Square no. 6, sector 1, Bucharest, code 010374, Romania, Phones: +40 21.319.19.00, +40 21.319.19.01, cosminlia@hotmail.com

How to cite: Minea V., Stan C., Florescu G.-D., Lianu C. & Lianu C. (2023). "Automated (Semantics Driven) Data Retrieval from Fiscal Documents: A Comprehensive Approach". *Annals of Spiru Haret University. Economic Series*, *23*(4), 327-342, doi: https://doi.org/10.26458/23416

Abstract

The importance of paper documents in regular business flow cannot be underestimated. They are an important part of the business domain increasingly digital landscape, complementing digital solutions by providing a plus of transparency, reliability and security. Making prompt decisions in the business world requires fast access to relevant and up-to-date data, and working with paper-based documents is very inefficient. Digitization of documents is ubiquitous, and digital document management systems (DMS) play an important role in fields like science, business or health. In the business domain, Enterprise Resource Planning (ERP) systems represent an entire ecosystem of solutions, meant to address every aspect of the business process, in a unified approach. An important aspect of successful ERP implementations

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

is related to the integration of DMS into the ERP. Enabling automated retrieval of data from all kinds of fiscal paper documents into the ERP is the next logical step. In this paper, we provide a hands-on approach for the task of automated text retrieval from fiscal documents. The novelty of our work resides in the manner in which we addressed the semantics of the retrieved data, such that the system associates meaning to the retrieved text elements, at the same time easing the processing of future documents. The solution is presented in a generic form, with a thorough discussion of the technological aspects. It is further implemented in the ERP system. We present and discuss experimental results, finally drawing conclusions and providing several ideas to further develop our work.

Keywords: *automated data retrieval, semantics-driven, fiscal documents, comprehensive approach, data extraction, semantic technology, information retrieval, document analysis, machine learning, financial data, tax documents, semantic web, natural language processing, computational linguistics, intelligent information retrieval, data integration, taxonomy, document processing, information extraction, semantic annotation*

JEL Classification: C88, D83, E62, G38, H83, O33

Introduction

Fiscal documents are documents utilized to register taxes, sales, purchases, transfer of ownership of goods or assets. Possessing legal validity, they constitute vital components in the audit and financial review processes. For regulation purposes, fiscal documents are required to be stored in physical format, in order to have them reviewed, in case of necessity [1]. Paper documents are a pillar of business domain, due to their tangibility and sense of permanence and advantages with respect to legal value, their potential for traceability of financial activities. The business domain relies heavily on the use of paper documents, such as contracts, invoices or receipts, despite the digital transformation era. Physical documents cannot yet be eliminated from the business flow: traditional industries, but not only them, rely on the use of paper documents because of their familiarity and the trust they still instil in people, but also because paper documents may serve parties in cases of regulatory compliance, basic accounting tasks and even for historical record keeping. Even though digital business solution prevail, paper documents

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

may, finally, serve as a dependable emergency backup of the businesses' information.

In today's modern world, the volume of information dealt in the business domain, and not only, is huge. The storage space required to store physical documents is a challenge for business owners, also raising security concerns due to their vulnerability to damage or loss. Searching data in a physical archive of paper documents may prove to be a challenging, time consuming task, while exposing such documents to various environmental conditions may contribute to their fast deterioration. Information retrieval from documents is essential for efficient collaboration, such that digitization of documents is a prevalent task across various business fields [8]. For example, in Brazil, the implementation of electronic invoices was intended to provide improved control over the tax assessment, and it led, as a side effect, to a significant increase in the collection of taxes [15].

Text extracted from paper documents is introduced into digital document management systems (DMS), where information can be efficiently found and managed, and easily searched. Digital DMS are highly scalable, handling large volumes of documents with ease, and they profit from all kinds of technological advances to ensure timely and reliable backups. DMS facilitate aggregation of data and collaboration among members of a company, allowing them to simultaneously work on the same documents and communicate by means of specific annotations on the text, while ensuring increased security by means of tracking access and changes to the digital documents, or by specific security policies that limit access to sensitive data.

Enterprise Resource Planning (ERP) systems represent a modern approach to managing and controlling business activities in a unified manner. Every aspect of the business activity such as planning, distribution, production and all business specific information is treated integratively in ERP systems. These systems are typically built as modular information systems, enabling the various modules to evolve and be maintained independently. DMS are an important module within well-designed ERP systems, handling all the information available in the system.

The quality of information stored in the ERP ecosystem is crucial for driving intelligent decision-making processes. Data within the ERP may originate from various sources, collected automatically from digital sources or manually from paper-based documents. Big Data Analytics (BDA) techniques work in conjunction with ERP systems to gather, analyse, and visualize data, to informing decision-making processes [3].

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

The importance of data stored in DMS is paramount, since important business decisions are based on business intelligence techniques that involve data analytics [2]. Manual data entry is a highly time-consuming process, heavily dependent on the knowledge and attention of individuals performing the task. Automating this process enhances efficiency, reduces erroneous data, and enables the swift processing of large document volumes.

Various technological solutions have emerged in recent decades to address the challenge of automated data capture, with Optical Character Recognition (OCR) being a notable example. OCR involves automatically recognizing characters from scanned images [5]. While OCR process is not error free, as errors are common, complete elimination of manual intervention by human operators is challenging [6].

This paper introduces a hands-on approach to integrate OCR techniques for the automatic processing of financial paper documents, aiming to enhance the data collection process in an ERP. Our interest extends beyond retrieving text from the paper documents: we focus on meaningful data extraction, ensuring that the retrieved data is stored in the database within specific fields. Our emphasis lies on both error detection and correction, involving user intervention during the data extraction process. We leverage the OCR capabilities implemented in Microsoft Azure, specifically Document Intelligence, and we seamlessly integrate them into the SeniorERP system. Experimental results exemplify and validate our approach.

In the following, we review specific works in the literature, related to the topic of our research. Then, we present in detail the research method, providing details on the integration process of Microsoft Azure's Document Intelligence component within the SeniorERP software system. The subsequent section focuses on validating the proposed methodology and presenting the experimental results. We conclude the paper by discussing the advantages and limitations of the proposed approach. Furthermore, we outline several directions for future research.

Related Work

ERP implementation is an endeavour of mature IT companies, both global players such as Oracle or SAS, or local IT firms such as Senior Software[1] or MentorSoft[2]. Each country has very specific regulations [14], making it a difficult task to use directly a generic ERP system out-of-the-box. This enables custom-made solutions, that are adapted to the particularities of a certain market, to have an

[1] https://www.seniorerp.ro/en/
[2] https://portal.winmentor.ro/winmentor/distributie/

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

advantage over general purpose ERP systems [9]. In Romania, the economic growth witnessed by the country since joining the EU has influenced greatly the offerings of IT companies with respect to software for management of all the aspects of the business, such as distribution, manufacturing, services, and retail activities.

In an ERP system, a vital component involves the data warehouse of the system, where data is stored securely, in such a manner that it allows easy access and search. Nowadays, the volume of data is so large that Big Data techniques are employed regularly in the business filed. Big Data Analytics have a significant impact on the development of management strategies and activities [3]. Integrating BDA in the digitalized healthcare supply chain in [4] they obtained improved efficiency and a strict control over the process of manufacturing and delivering of medicines [4].

At the core of BDA lies the data that the analysts use, data that is usually automatically fetched from the various modules of an ERP system. Data from paper documents is usually manually introduced in the system, but great advantages are obtained from the automation of this task [10]. Optical Character Recognition is the task of converting handwritten or printed documents into digital data, by automatically identifying and recognizing the text. A systematic review of OCR capabilities available is documented in [7]. Automatic processing of documents achieves a low degree of intrusiveness and significant reduction of operational costs [9]. From the point of view of implementation of Industry 4.0 processes, "robotic process automation" entails text recognition by means of OCR based software, followed by automatic data extraction and automatic document classification, when such a process is possible [10].

Following OCR recognition of characters, the retrieved text is further processed. Post-processing involves the detection of errors [11], which may involve segmentation errors or even recognition errors. Further, error correction is attempted, and this process usually involves the generation of plausible candidates to replace the erroneous words [11]. Other more sophisticated approaches to error correction involve the use of a lexicon of accepted words [12]. When dealing with documents that have a structured format, error correction may address pairing information from the recognized text and the expected structure of the document [13].

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Various platforms offer powerful OCR algorithm implementation as services [16], such as Azure AI Document Intelligence[3], Google Document AI[4], Amazon Textract[5], Base64.ai[6], Clova OCR[7], AbbyCloud OCR[8], Parashift.IO[9] or IBM Cloud[10], Rossum AI[11]. Among these, Azure's Document Intelligence (ADI) features recommend it for high precision OCR tasks [17], in contexts such as integration with ERP system [18]. From our point of view, ADI is the preferred choice also because it offers excellent support for the Romanian language.

After extraction of text from paper documents, it must be associated with semantics, in order to truly be valuable for search and analysis. Information Retrieval (IR) is a task related to OCR, but with the focus on the meaning of the data retrieved from documents [19]. In [20], an OCR Search Engine is described in order to extract meaningful information from digitized books. In [21], a library of OCR annotations for industry documents is presented, in order to offer a baseline comparison of various commercial tools. Three methods for "hand-marked semantic text recognition" are presented in [22], where images from scanned paper documents are processed in order to extract semantic text. Several quantitative measures are introduced in [23], in order to evaluate dataset specific biases in automated document annotation tasks.

Proposed approach

The creation of semantic association between automatically recognized text from paper documents and meaningful information from the specific domain of the text is a task of great interest in the context of the digitization of documents in various fields. In our approach, we build on Azure Document Intelligence (ADI) and provide an original approach for the semantic annotation of automatically OCR documents. ADI offers a General Document Model (GDM) which extract data stored in documents, such as text, tables, key-value pairs and text. We embed this

[3] https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence
[4] https://cloud.google.com/docu ment-ai/docs/overview?hl=en_US
[5] https://docs.aws.amazon.com/textract/latest/dg/what-is.html
[6] https://documenter.getpostman.com/view/10132588/SWT5hfdz
[7] https://www.ncloud.com/product/aiService/ocr
[8] https://www.abbyy.com/vantage/ocr-skill/features/
[9] https://docs.parashift.io/#intro
[10] https://cloud.ibm.com/docs/discovery-data
[11] https://elis.rossum.ai/api/docs/#getting-started

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

and propose specific workflow in order to structure and organize information, and link it to business concepts integrated in the ERP system, such that the information is fructified within the ERP system.

Methodology

The transformation of a mage-type Document into OCR Data is described in the following. We implemented this step as a web application, such that it is offered as a service. In Figure 1, we present the dynamic view of the OCR application behaviour, by means of a sequence diagram. It depicts the interaction between the user, the OCR module of the ERP application and Microsoft Azure Document Intelligence.

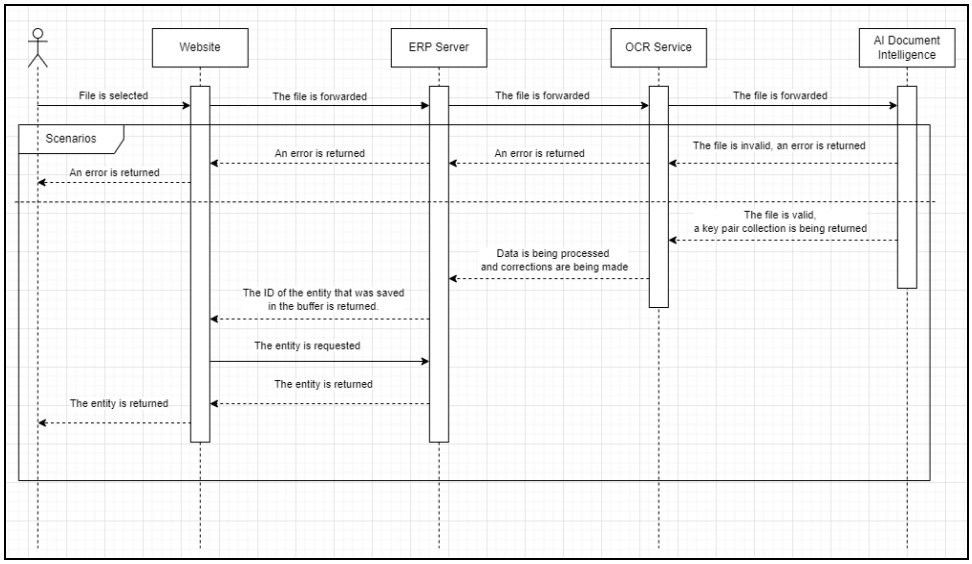

Figure 1: Sequence diagram of the OCR flow

The process is described in more detail in the following (see Figure 2). As a first step, the user accesses the website, where they have the option to upload a PDF or image document and select the document model they want to use. There exist several predefined template documents for some common types of specific invoices (such as those from some common companies in the Romanian market). The document is transformed into a vector of bytes and, along with information

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

about the model, is transmitted to the ERP server. The ERP server forwards the data to the OCR service, hosted in the Microsoft Azure Cloud. The OCR service calls AI Document Intelligence to transform the document into OCR data. At this stage, there exist two possible scenarios, from our point of view.

In the first scenario, AI Document Intelligence returns that the file is not valid or does not correspond to a specific model and the optical character recognition process failed. The OCR service receives the error and passes it on to the ERP server. The ERP server further sends the error to the website. The website displays a user-friendly message informing the user of the error.

In the second scenario, AI Document Intelligence signals that the transmitted file is valid and returns a key-value collection with the results obtained from the OCR process. The OCR service receives the information, processes the data, and makes corrections where necessary, then forwards the information to the ERP server. The ERP server saves the data in a Buffer table and returns the ID of the saved entity to the website. The website requests the data for the entity with the received ID. The ERP server returns the requested entity. The website displays the user with the data resulting from the OCR process.

In order to assimilate the OCR Data into a valid ERP Document, the OCR document is sent for conversion. During the conversion process, it is checked whether the business partner (individual or legal entity on the OCR document) already exists in the ERP database. Mapping is done by searching for a partner in ERP with the same code or name. If no entity with these characteristics is identified, the partner is marked as unmapped. If there is an entity in ERP that matches the characteristics, its data is set in the new ERP document being constructed. It is checked if there is a county in the database that matches the one in the OCR document. Similarly, if it does not exist, the OCR document is marked as unmapped for the county. If it does exist, its data is set in the ERP document. The process is repeated for other document entities: City, Document Currency, General VAT Rate, etc.

A detail is then read from the tabular area of the document. By detail, we mean information related to a product on the document, including code, name, unit of measure, quantity, price, net value, gross value etc. It is searched in the database if a product with the same barcode as the one in the OCR document exists. If found, it proceeds to setting the data for a new ERP document detail. It is searched in the database if there is a product with any alternative name or code set to match the name or code of the product in the OCR document detail. If found, it proceeds to setting the data for a new ERP document detail.

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Issue 4/2023

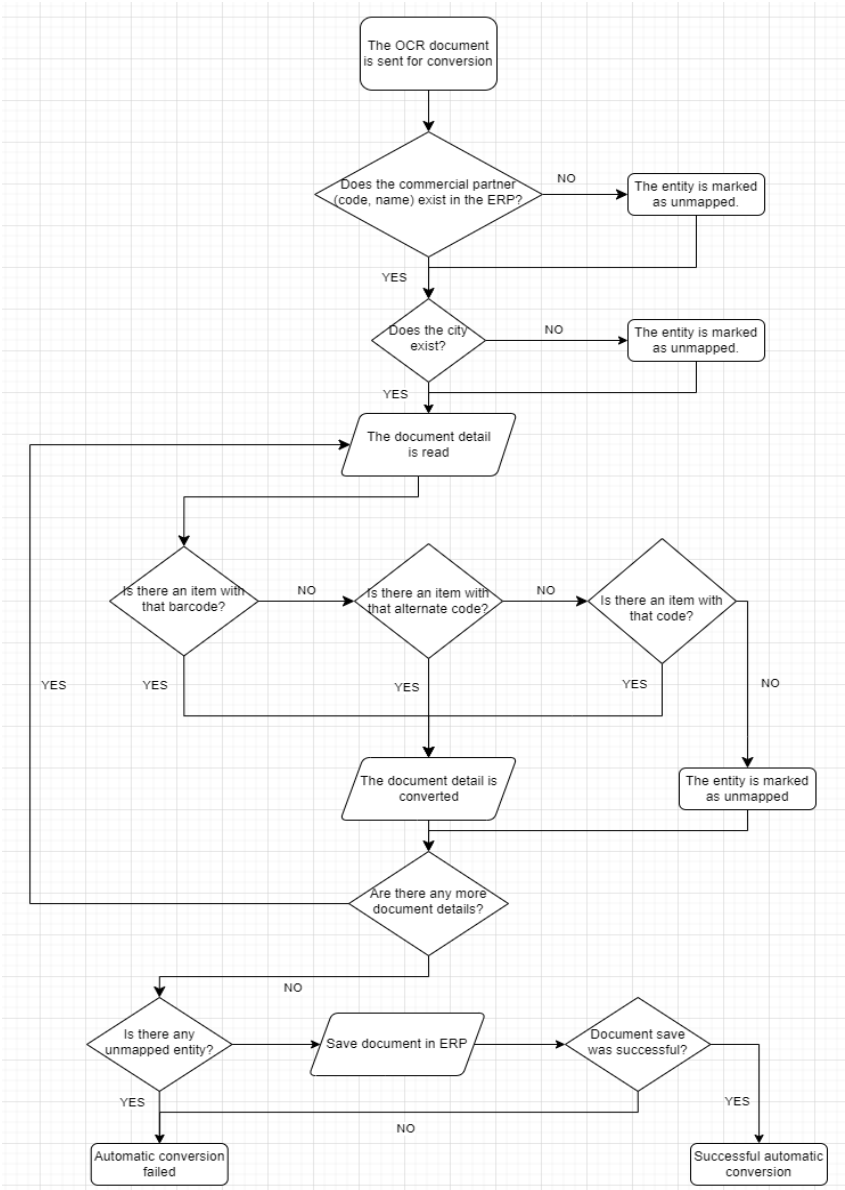

Figure 2: OCR Application Usage Flow: Transformation of Image-Type Document into OCR Data, Transformation of OCR Data into ERP Document.

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

An ERP database product can have alternative codes and names set (often being the names customers and suppliers use for the same product) to help with future mapping. If automatic mapping fails, the user can manually select a product, in which case the values for code and name that appear in the OCR are automatically set to ensure automatic mapping in the future. The data of the identified product obtained from the database is set in a new detail of the ERP document being constructed. It is searched in the database if there is a product with a name or code that matches the name or code of the product in the OCR document detail. If not found, the OCR document detail is marked as unmapped. It is checked if there are other details on the OCR document that have not been read. If there are, it returns to the stage of reading another detail from the document. Finally, it is checked if there are entities in the OCR document that could not be mapped. If there are, the conversion of the OCR document to the ERP document is marked as failed. If not, an attempt is made to save the constructed document in ERP. If the saving in ERP is successful, the conversion of the OCR document to the ERP document is marked as successful. Otherwise, it is marked as failed.

The user has the option to intervene and manually select entities that could not be identified automatically by the system. Within some of these, associations are made in the system so that for a future conversion with similar values, the identification will be automatic.

Performance Measures

For the evaluation of the performance in the OCR tasks, we used Word Error Rate (WER) and Entity Error Rate (EER). These are widely used in similar tasks in the literature.

The WER is defined as WER = (S + D + I)/N, where

- S denotes the number of words incorrectly identified
  For example: the word "*Calea*" may be misidentified as "*Caiea*" because the letter "l" is wrongfully identified as "i"
- D – is the number of missed words
  For example: "*Furnizor SC IMPEX SRL"* may be identified as "*Furnizor SC*"
- I - the number of words erroneously added (these words do not exist in the real text, but are returned as recognized by the system)
  For example: "*Piulita S29"* may be recognized as "*Piu lia S29"*

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

The number of correct words is denoted as C, and the total number of words is N. It results that

$$N = S+D+C.$$

For example, if we consider the original and the recognized texts to be as in Table 1 below:

Table 1: An example of original text and automatically recognized text.

"*Furnizor SC IMPEX SRL*	"*Furnizor SC*
Adresa: Calea Ferentari nr. 8	*Adresa: Caiea Ferentari nr. 8*
Articole:	*Articole:*
Piulita S29 3 bucati x 0.3 lei"	*Piu lita S29 3 bucati x 0.3 lei"*

It results that:

S = 1 (Caiea)

D = 2 (IMPEX SRL)

I = 2 (Piu lita)

C = 13

and the WER = 6/16 = 0.375, hence a percentual error of 37.5%.
The Entity Error Rate (EER) represents the percentage of misidentified entities in a document. It is computed as:

$$EER = Et / Nt$$

where Et – is the number of misclassified entities, and Nt - is the total number of entities in the document.

Experimental results

The system was thoroughly tested, on fiscal documents specific to the Romanian market. Such a document is presented in Figure 3. We tested the system on other documents, where various fields were missing, or document where the text was hardly visible or hardly readable.

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Issue 4/2023

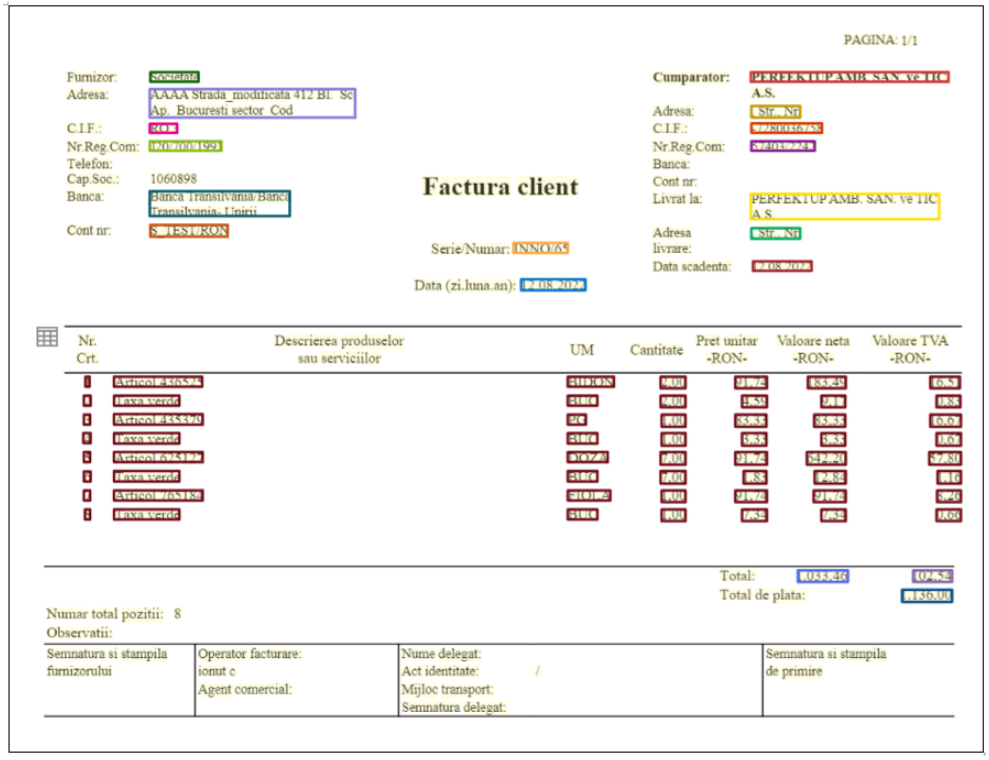

Figure 3: Automatic labels in a document, placed by Document Intelligence.

In a battery of 10 tests, involving variations of the standard fiscal document of above, we annotated 20 labels and, on top of these, the values for each item in the fiscal document. The accuracy of the annotations identified is presented below (Figure 4).

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Issue 4/2023

Field Name	Accuracy
ProviderName	99.50 %
Provider Address	90.00 %
ProviderCIF	99.50 %
InvoiceNumber	90.00 %
InvoiceDate	90.00 %
ClientName	90.00 %
ClientAddress	70.00 %
ClientCIF	99.50 %
InvoiceDueDate	99.50 %
articles-details	99.50 %

DeliveryAddress	90.00 %
DeliverTo	70.00 %
ProviderTRN	99.50 %
ClientTRN	99.50 %
ProviderBank	90.00 %
ProviderBankAccount	99.50 %
ClientBank	90.00 %
ClientBankAccount	90.00 %
NetTotal	99.50 %
TVATotal	90.00 %
TotalToPay	99.50 %

A summary of the results for the 10 tested fiscal documents is presented below:

File	S	D	I	C	N	WER
48.pdf	0	0	1	55	55	0.018
29.pdf	0	0	0	79	79	0
25.pdf	0	0	0	66	66	0
34.pdf	0	1	0	96	97	0.01
20.pdf	0	0	0	68	68	0
60.pdf	0	0	0	84	84	0
22.pdf	0	0	0	66	66	0
54.pdf	0	0	4	69	69	0.058
47.pdf	0	0	0	106	106	0
37.pdf	0	0	2	90	90	0.022
					Average	0.0108

Annals of Spiru Haret University.
Economic Series
Since 2000
ISSN:2393-1795 ISSN-L:2068-6900

Conclusions

This paper tackled the task of implementing OCR text recognition for fiscal documents as a service provided in a complex ERP system. The OCR process is realized by the integration of Microsoft Azure Document Intelligence system. Our approach involved the extraction of text and the automatic annotation of the data such that it is stored as an ERP document, in the system's data warehouse. To this end, we devised a workflow that proceeds in well-designed stages in order to treat all the data received from the OCR service. In this respect, we offer ERP system users the possibility to accept automatic labelling of the retrieved entities, or to assign new labels, which are memorized and will be made available in future processing of documents. Digitization of paper fiscal documents facilitates quicker analysis of financial data, allowing for faster processing and easier interpretation. It reduces the time and effort required for data entry, minimizing errors and streamlining fiscal document processing. It serves purposes of compliance and auditing, because it helps with fast and accurate record keeping, with small costs and reduced need of manual labour. Digitized fiscal documents that are further annotated, become meaningful resources of information, easily searchable. We provided a solution for a pivotal problem in the process of digital transformation of business processes, by using OCR for fiscal paper documents.

Funding: This research was funded by the European Regional Development Fund, based on financing contract no. 27/221_ap3/22.07.2022, concluded with the Romanian Digitalization Authority, as an Intermediary Body for the Operational Competitiveness Program, project “Innovative platform for streamlining the activity of SMEs in Romania”, code MySMIS 142621.

Acknowledgements: We are grateful for the guidance of Mrs. Elena Bautu, from the "Ovidius" University Constanta, whose comments and insights significantly improved our paper.

Bibliography

[1] Accounting Records: Definition, What They Include, and Types, https://www.investopedia.com/terms/a/accounting-records.asp
[2] Ragazou, Konstantina, Ioannis Passas, Alexandros Garefalakis, Emilios Galariotis, and Constantin Zopounidis. 2023. "Big Data Analytics Applications in Information Management Driving Operational Efficiencies and Decision-Making: Mapping the Field

of Knowledge with Bibliometric Analysis Using R" Big Data and Cognitive Computing 7, no. 1: 13. https://doi.org/10.3390/bdcc7010013

[3] Abdelhalim, A.M. (2023), "How management accounting practices integrate with big data analytics and its impact on corporate sustainability", Journal of Financial Reporting and Accounting, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JFRA-01-2023-0053

[4] Surajit Bag, Pavitra Dhamija, Rajesh Kumar Singh, Muhammad Sabbir Rahman, V. Raja Sreedharan, "Big data analytics and artificial intelligence technologies based collaborative platform empowering absorptive capacity in health care supply chain: An empirical study", Journal of Business Research, Volume 154, 2023, 113315, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2022.113315

[5] K. M. Yindumathi, S. S. Chaudhari, and R. Aparna, “Structured data extraction using machine learning from image of unstructured bills/invoices,” in Smart Computing Techniques and Applications, S. C. Satapathy, V. Bhateja, M. N. Favorskaya, and T. Adilakshmi, Eds. Singapore: Springer Singapore, 2021. ISBN 978-981-16-1502-3 pp. 129–140. [Page 2.]

[6] Irimia, Cosmin, Florin Harbuzariu, Ionut Hazi, and Adrian Iftene. "Official Document Identification and Data Extraction using Templates and OCR." Procedia Computer Science 207 (2022): 1571-1580.

[7] Mittal, Rishabh, and Anchal Garg. "Text extraction using OCR: a systematic review." In 2020 second international conference on inventive research in computing applications (ICIRCA), pp. 357-362. IEEE, 2020.

[8] Ma, Ke, Sagnik Das, Zhixin Shu, and Dimitris Samaras. "Learning from documents in the wild to improve document unwarping." In ACM SIGGRAPH 2022 Conference Proceedings, pp. 1-9. 2022.

[9] Iryna Lukyanova; Abubaker Haddud; Anshuman Khare. "Types of ERP Systems and Their Impacts on the Supply Chains in the Humanitarian and Private Sectors." Sustainability 14, no. 20, 2022: 13054. https://www.mdpi.com/2071-1050/14/20/13054

[10] Jorge Ribeiro; Rui Lima; Tiago Eckhardt, and Sara Paiva. "Robotic process automation and artificial intelligence in industry 4.0–a literature review." Procedia Computer Science 181, 2021; pp. 51-58.

[11] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, “Survey of postOCR processing approaches,” ACM computing surveys, vol. 54, no. 6, pp. 1–37, 2021. doi: 10.1145/3453476

[12] Rijhwani, Shruti, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. "Lexically aware semi-supervised learning for OCR post-correction." Transactions of the Association for Computational Linguistics 9 (2021): 1285-1302.

[13] N. Kamaleson, D. Chu, and F. E. B. Otero, “Automatic information extraction from electronic documents using machine learning,” in Artificial Intelligence XXXVIII, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 183–194. ISBN 9783030910990

[14] Genifera Claudia Bănică. 2022. Results of the Fiscal Control Activity in Romania and Other Europeans States. UTMS Journal of Economics 13(1): 29–42.
[15] Vieira, Patrícia Araújo, Daiana Paula Pimenta, Alethéia Ferreira da Cruz, and Eliane Moreira Sá de Souza. "Effects of the electronic invoice program on the increase of state collection." Revista de Administração Pública 53 (2019): 481-491, https://doi.org/10.1590/0034-761220170077
[16] Cutting, Graham A., and Anne-Françoise Cutting-Decelle. "Intelligent Document Processing--Methods and Tools in the real world." arXiv preprint arXiv:2112.14070 (2021).
[17] Salvaris, Mathew, Danielle Dean, and Wee Hyong Tok. "Deep learning with Azure." Building and Deploying Artificial Intelligence Solutions on Microsoft AI Platform, Apress (2018).
[18] Price, E., Masood, A. and Aroraa, G., 2021. Hands-on Azure Cognitive Services: Applying AI and Machine Learning for Richer Applications. Apress LP.
[19] Callan, Jamie, Paul Kantor, and David Grossman. "Information retrieval and OCR: from converting content to grasping meaning." In ACM SIGIR Forum, vol. 36, no. 2, pp. 58-61. New York, NY, USA: ACM, 2002.
[20] Gupta, Riya, and C. V. Jawahar. "Information Retrieval from the Digitized Books." arXiv preprint arXiv:2212.00999 (2022).
[21] Biten, A.F., Tito, R., Gomez, L., Valveny, E., Karatzas, D. (2023). OCR-IDL: OCR Annotations for Industry Document Library Dataset. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds) Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol 13804. Springer, Cham. https://doi.org/10.1007/978-3-031-25069-9_16
[22] Suh, S., Lee, G., Gil, D. et al. Automated hand-marked semantic text recognition from photographs. Sci Rep 13, 14240 (2023). https://doi.org/10.1038/s41598-023-41489-4
[23] Ayushi Dutta, Yashaswi Verma, C. V. Jawahar, "Automatic image annotation: the quirks and what works", Multimed Tools Appl (2018), 77:31991-32011, DOI 10.1007/s11042-018-6247-3