Using RPA for data generation using OCR platforms in Mediterranean University of Albania

Gerild Qordja,
PhDp, Mediterranean University of Albania, Faculty Of Informatics, Department of Information Technology Tirana, Albania
E-mail address: gerildqordja@umsh.edu.al

Abstract
The increase in the amount of data today has led to the use of computer applications in order to manage processes precisely. Robotic process automation (RPA), also known as software robotics, uses automation technologies to mimic back-office tasks of human workers, such as extracting data, filling in forms, moving files, et cetera. Optical character recognition (OCR) is sometimes referred to as text recognition. An OCR program extracts and repurposes data from scanned documents, camera images and image-only pdfs. OCR systems use a combination of hardware and software to convert physical, printed documents into machine-readable text. Hardware such as an optical scanner or specialized circuit board copies or reads text then, software typically handles the advanced processing. Process Automation in Azure Automation allows you to automate frequent, time-consuming, and error-prone management tasks. This service helps you focus on work that adds business value. In this paper, I will use the above-mentioned technologies to realize the automatic data generation process for the construction of an online library. In addition, the level of data accuracy will be studied in the automation of data generation from pdf files to mySql. The application will be built in front end html and back end php programming language and mySql database. These tests will be done by inserting more than 17000 books in pdf format.

Keywords: Microsoft Azure, Robotic Process Automation (RPA), Optical Character Recognition (OCR), MySql, Html.

1. Introduction
Nowadays, technology is coming into use more and more. In this paper, I will provide optimal solutions for the process of unstructured data into structured data using OCR applications.

HTML is the predominant markup language for webpages. It uses tags to create structured documents via semantics for text—such as headings, paragraphs, and lists—as well as for links and other elements. HTML also lets authors embed images and objects in pages and can create interactive forms [1].

As a visual database design tool for the MySQL database system, MySQL Workbench combines SQL development, database design, construction, and maintenance into a single integrated development environment [2].

Automating operational tasks is critical for streamlining infrastructure management, both on premises and in the cloud. Microsoft Azure Automation comes with capabilities that help administrators automate their cloud-based, operational, repetitive tasks [3].

The open source application that became Calibre was created by Kovid Goyal in 2006 under the name "libprs500". As it became popular a name change was suggested, and "Calibre" was chosen by Goyal's \Vifc Krittika [4].

2. Literature Review

Information systems for University administration during the pandemic have resulted in China being extremely efficient [5]. These systems have been implemented in Chinese universities to automate the control of current systems.

The automation of libraries in universities has been very efficient in reducing the order of services to a minimum [6]. This process saves time, minimizes errors, increases the efficiency of the process compared to the traditional process.

Various private companies and, of course, state institutions have started implementing applications to automate various business processes since early times. In the times we are talking about, investing in Information and Communication Technology resources is no longer a choice option but an obligation for businesses [7]. In recent years, many public and private organizations have changed the way of thinking about the solutions of their business processes to improve the quality of the services provided, achieving a better efficiency [8].

Since the 1993s, the concept of automated business processes appeared, which with the development of technology have become indispensable in the stable and real-time management of the progress of a business. Business Process Reengineering is "the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical contemporary performance measures such as cost, quality, service and speed".

Workflow Management System (WfMS) is another example of technology that enables improved process performance in a collaborative network environment. A Workflow Management System (WfMS) enables process automation through the integration, coordination, and communication of human and automated tasks of a business process.

The discipline of business process management (Business Process Management) investigates methods and techniques to organize business processes in an efficient and effective way.

Recent advances in the field of Artificial Intelligence, Machine Learning, Cryptography and distributed systems have provided the foundations for new technologies, including robotic process automation, chatbots, machines with self-driving, smart objects, blockchains and the Internet of Things. [9]

Several recent papers discuss the implications of the emergence of these technologies for BPM. These technologies are likely to influence the way organizations design and execute business processes in the future. However, it is not clear in what specific way they will change BPM. [10].

It is likely that new Human Resources management systems will realize the possibility of managing personnel costs and mapping business processes for each department. Such an

advantage will help automate the enterprise's unified corporate system and bring it to a new level by reducing costs and increasing competition [11, 12].

3. The case of the University

The Mediterranean University of Albania is one of the private higher education institutions in the Republic of Albania accredited at 4 (four) levels of study, professional diploma, bachelor's, master's and doctorate.

As a higher education institution in Albania, as well as private companies and other institutions in various sectors, the Mediterranean University of Albania has to manage various sensitive internal and external data from time to time in real time and limited.

The Mediterranean University of Albania had about 17,000 books in pdf format, which needed to be in the web application. It was impossible for all the unstructured data of the books, such as the title, author, etc. to be manually converted into data formats. structured. This is because it was required that each book be opened manually and the data noted above be stored in the database.

To create the suggested application, which is presented in point 4 of this paper, Neatbeans version 8.02 technologies were used as a compiler, MySql Workbeanch as a database, and Caliber as an OCR application to generate structured data. HTML, CSS, PHP are used as programming languages.

Before this application was active, the physical bookstore was active. The work process of receiving and returning books was completely manual. All management of this process was documented through management forms. Through this application, it will be possible to digitize and search these books by all university users who want to have digital books.

4. Suggested web application

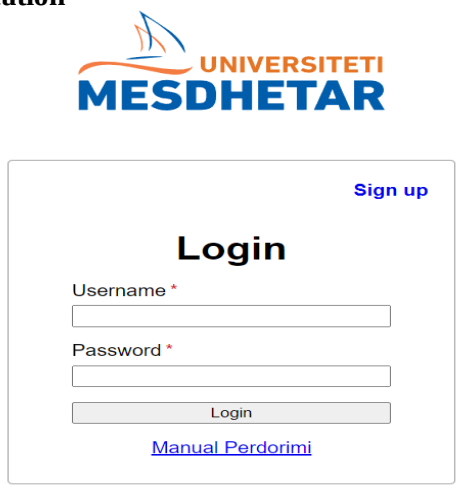


Fig. 1. Login interface
Source: Author own work

In the login interface in the figure 1, the user can interact with the interface by logging into the system through the Username and Password that he determines when he registers in this web application. Also on this interface, a user manual has been applied. If the user does not have account you can create one by clicking on Sign up.

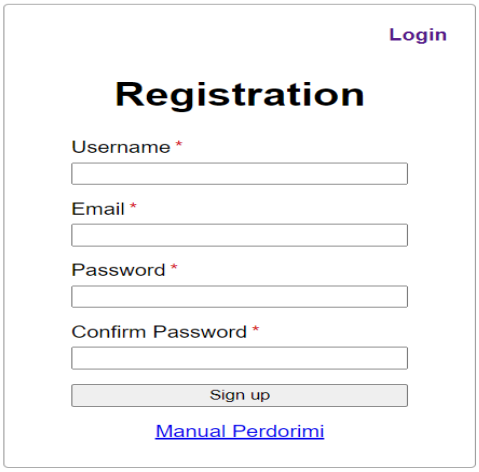


Fig. 2. Registration interface
Source: Author own work

Users not registered in the interface of figure 2 can create an account by inserting Username, Email, Password in the database. If the user is found in the database with a username or email, the application generates an error.

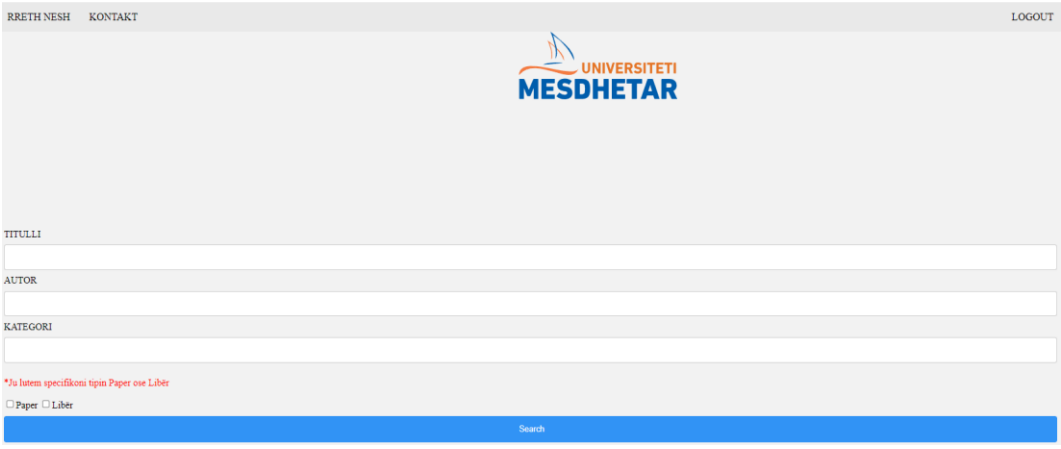


Fig. 3. The main book or paper search interface
Source: Author own work

The interface in Figure 3 shows the way and the filters used to filter books by title, author, and category. You can search for a book according to one of the campaigns or according to all three.At the same time, you must choose the type of pdf you are looking for, paper or book.

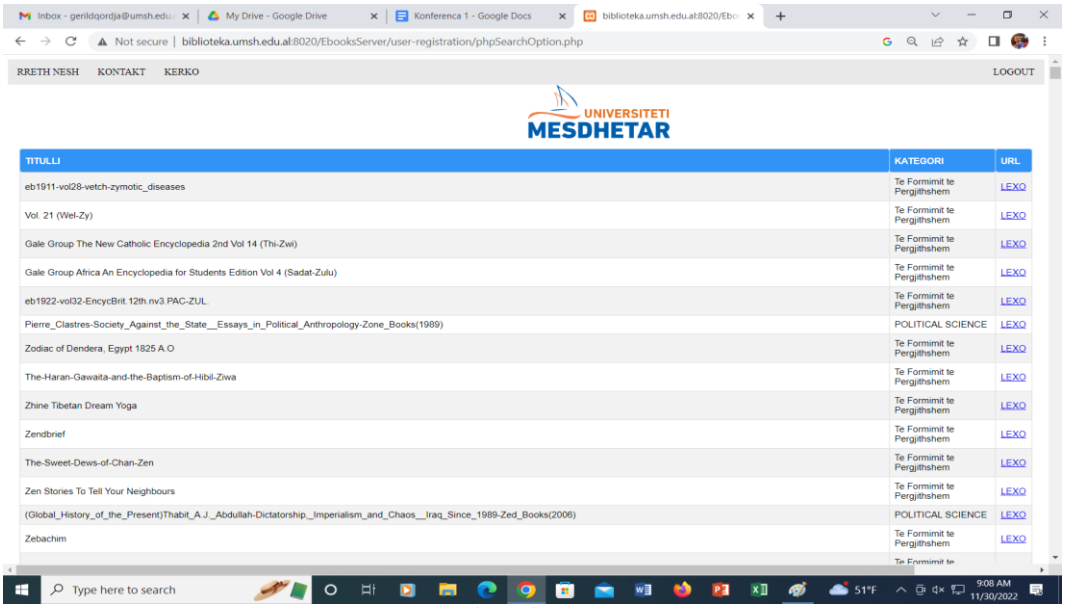


Fig. 4. The main interface after clicking the search event
Source: Author own work

After the Search Book functionality is called at the moment when we have not selected any of the filtering fields, all books will be displayed in total.By clicking on the LEXO link, we will be able to open the pdf with the respective books.

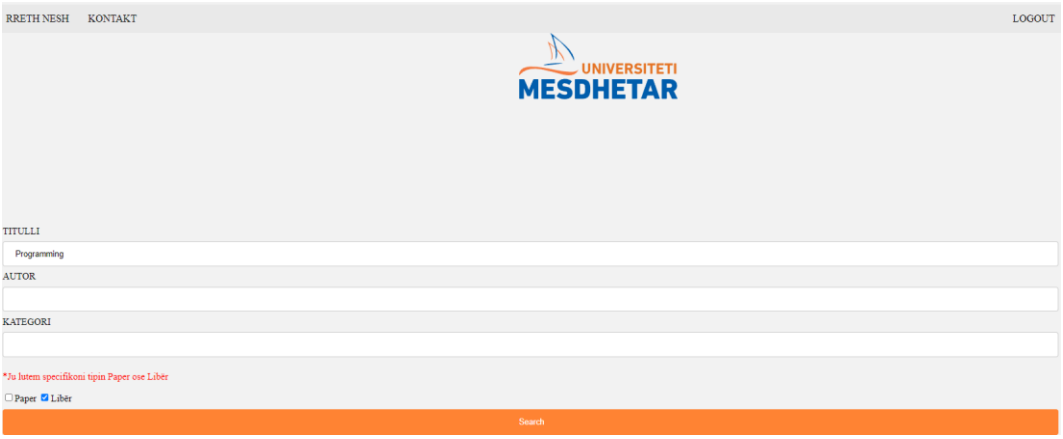
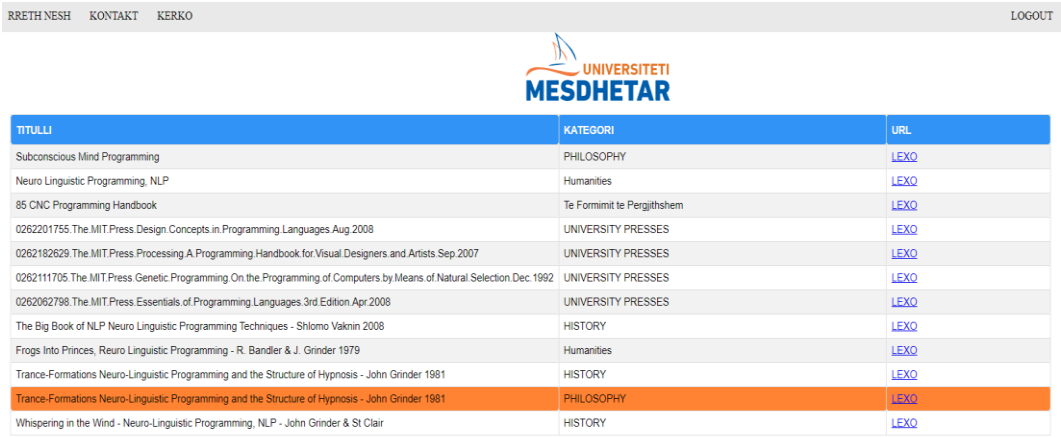


Fig. 5. Main interface with filter by title
Source: Author own work

The page in figure 5 shows how to search for books by title.



TITULLI	KATEGORI	URL
Subconscious Mind Programming	PHILOSOPHY	LEXO
Neuro Linguistic Programming, NLP	Humanities	LEXO
85 CNC Programming Handbook	Te Formimit te Pergjithshem	LEXO
0262201755.The.MIT.Press.Design.Concepts.in.Programming.Languages.Aug.2008	UNIVERSITY PRESSES	LEXO
0262182629.The.MIT.Press.Processing.A.Programming.Handbook.for.Visual.Designers.and.Artists.Sep.2007	UNIVERSITY PRESSES	LEXO
0262111705.The.MIT.Press.Genetic.Programming.On.the.Programming.of.Computers.by.Means.of.Natural.Selection.Dec.1992	UNIVERSITY PRESSES	LEXO
0262062798.The.MIT.Press.Essentials.of.Programming.Languages.3rd.Edition.Apr.2008	UNIVERSITY PRESSES	LEXO
The Big Book of NLP Neuro Linguistic Programming Techniques - Shlomo Vaknin 2008	HISTORY	LEXO
Frogs Into Princes, Reuro Linguistic Programming - R. Bandler & J. Grinder 1979	Humanities	LEXO
Trance-Formations Neuro-Linguistic Programming and the Structure of Hypnosis - John Grinder 1981	HISTORY	LEXO
Trance-Formations Neuro-Linguistic Programming and the Structure of Hypnosis - John Grinder 1981	PHILOSOPHY	LEXO
Whispering in the Wind - Neuro-Linguistic Programming, NLP - John Grinder & St Clair	HISTORY	LEXO

Fig. 6. Main interface after clicking event search with filter by title
Source: Author own work

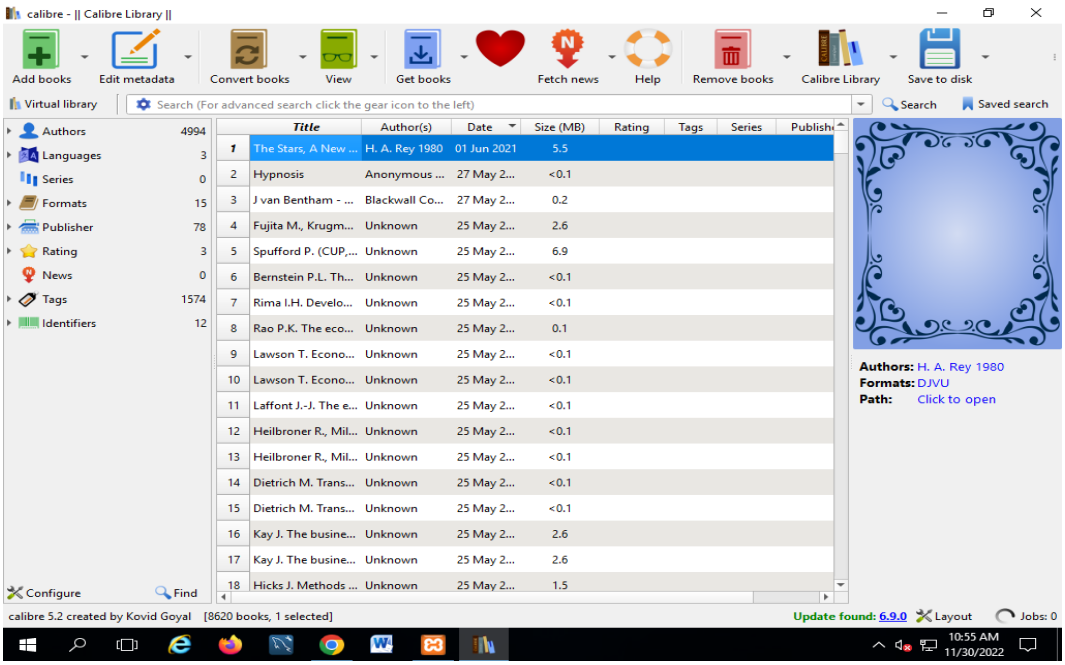


Fig. 7. Caliber main program interface
Source: Author own work

Figure 8 shows the Caliber application, which is used in this case as an OCR Application. In this application, the pdf books were loaded and then the data was generated which we will explain in figure 8.

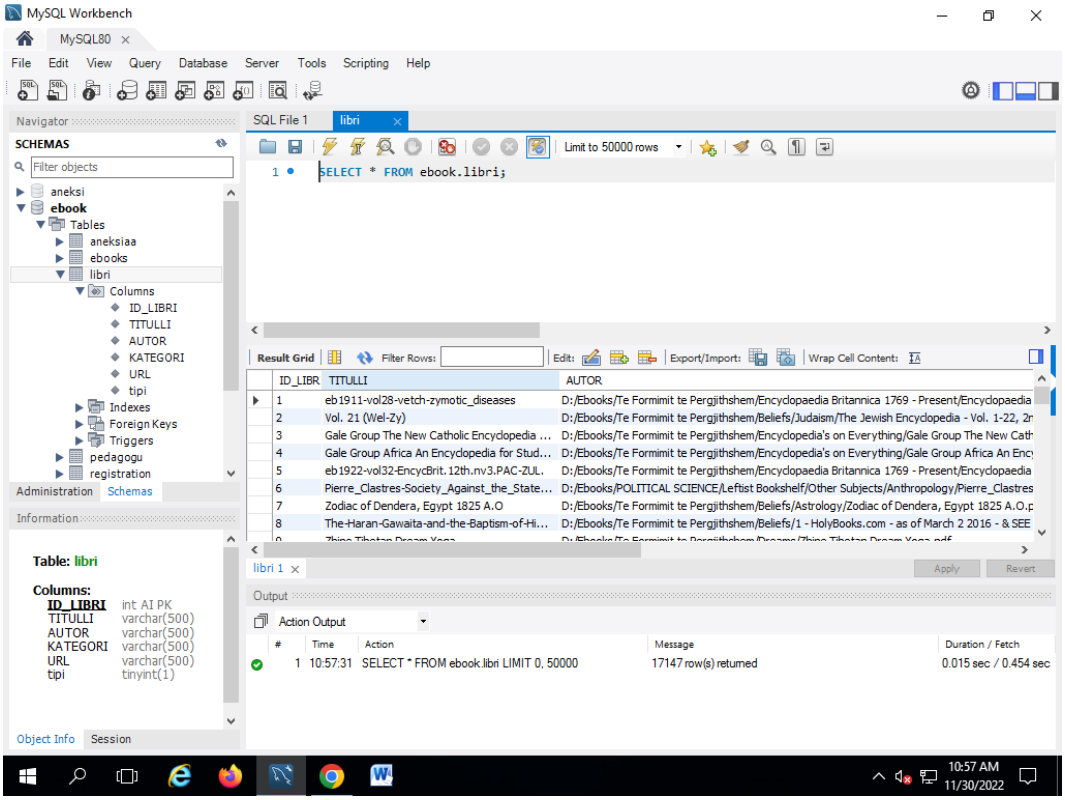


Fig. 8. The main SQL Workbench database interface
Source: Author own work

Figure 8 shows the database that is used to store in a structured way the data that will be used to access the book that will be searched. In this case, it is the ebook database and the table named book. The book table is stored in columns. its data as ID_Book which is unique, TITLE, AUTHOR, CATEGORY, URL, type.

The URL is automatically generated using the FILE SERVER service through RPA in Windows. While the type saves, if the pdf is liber or papper.

All books must be in the FILE of the NeatBeans work environment in order for the URL to be generated.

Conclusion

After applying the solution to the aforementioned problem of generating structured dates, we reached the following conclusions:

1. The OCR Caliber application can generate data up to 8000 only page one pdf materials in a generation event, more than that the structured data comes out with errors.
2. Application OCR systems have 100% accuracy in generating data compared to the input of unstructured data that is processed.

3. Using the Caliber program effectively generates data from unstructured to structured.
4. The structured data from the Caliber application can be inserted into the SQL Workbench database through a simple insert query.
5. The use of the suggested gene is used in the administration of books from the library, reducing time and increasing search performance.

References

[1] Seto, T., Nagafuji, T., Toyama, M. (1997), *Generating html sources with tfe enhanced sql*, In Proceedings of the 1997 ACM symposium on Applied computing, pp. 96-100.

[2] Daga, A., Dash, .D, Development Of An Internal Data Visualization Platform.

[3] Karthikeyan, S. A. (2017), *Azure automation using the ARM model: an in-depth guide to automation with Azure resource manager*, Apress.

[4] Jermey, J. (2014), *Calibre for ebook management*, ONLINE CURRENTS, 28(2), pp. 75-78.

[5] Li, R., & Chen, H. (2022), *Research on Automation Control of University Logistics Management System Based on Wireless Communication Network*, Wireless Communications and Mobile Computing.

[6] Tahil, S. K. (2022), *Library Automation: An Emerging Technology for State University and Colleges in Sulu Province*, Natural Sciences Engineering and Technology Journal, 2(1), pp. 85-89.

[7] Malenje, J. O., Otanga, D., Wadwoba, F. (2014), *Effective Business Process Automation through Process Reengineering: Case of Public a University in Kenya*, International Journal of Information and Communication Technology Research, 4(6), pp. 246-254.

[8] Holz, H. J., Applin, A., Haberman, B., Joyce, D., Purchase, H., & Reed, C. (2006), *Research Methods in Computing: What are they, and how should we teach them?*, In Working group reports on ITiCSE on Innovation and technology in computer science education, pp. 96-114.

[9] Mendling, J., Decker, G., Hull, R., Reijers, H. A., & Weber, I. (2018), *How do machine learning, robotic process automation, and blockchains affect the human factor in business process management?*, Communications of the Association for Information Systems, 43(1), p. 19.

[10] Holz, H. J., Applin, A., Haberman, B., Joyce, D., Purchase, H., & Reed, C. (2006), *Research Methods in Computing: What are they, and how should we teach them?*, In Working group reports on ITiCSE on Innovation and technology in computer science education, pp. 96-114.

[11] Velikorossov, V. V., Filin, S. A., Genkin, E. V., Maksimov, M. I., Krasilnikova, M. A., & Rakauskiyene, O. G. (2020), *HR systems as a new method for the automatization of business processes in organization*, In 2nd international conference on pedagogy, communication and sociology (ICPCS No. 2020), p. 415.

[12] Holz, H. J., Applin, A., Haberman, B., Joyce, D., Purchase, H., Reed, C. (2006), *Research Methods in Computing: What are they, and how should we teach them?*, In Working group reports on ITiCSE on Innovation and technology in computer science education, pp. 96-114.