BIS INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

Volume 2 2025 | DOI: 10.31603/bistycs.179
6th Borobudur International Symposium on Science and Technology (BIS-STE) 2024

UNIMMA PRESS
Universitas Muhammadiyah Magelang

OCR implementation in archiving system at Borobudur Subdistrict using regular expression and TextRank methods

Arif Zain[1*], Agus Setiawan[1], Dimas Sasongko[1]

[1] Universitas Muhammadiyah Magelang, Magelang 56553, Indonesia
[*]Corresponding author email: arif.zain502@gmail.com

Abstract

Borobudur Subdistrict is part of a government agency that is required to manage archives, including the registration of incoming letters. Over the past five years, approximately 10,000 incoming letters have been stored in the Borobudur Subdistrict, which must be registered and digitized to support implementing an Electronic-Based Government System. The efforts for registration and digitization have been carried out using conventional methods, which require a considerable amount of time. Therefore, this study aims to develop a system that can assist in the process of archive registration and digitization using Optical Character Recognition (OCR) techniques along with Regular Expression and TextRank methods. The system is designed to extract text from physical documents into digital text through OCR, automatically detecting required text patterns using the Regular Expression method and summarizing documents using the TextRank method. The study's results show a significant increase in efficiency, up to 300%, where the number of archives registered and digitized in one day increased from 20 to 80. This solution proves that implementing this system can significantly improve the speed of the process of archive management in the Borobudur Subdistrict and also provide an effective solution for the registration and digitization process.

Keywords

Digitization, Archive, OCR, Regular expression, TextRank

Published:
April 28, 2025

Selection and Peer-review under the responsibility of the 6th BIS-STE 2024 Committee

Introduction

Archives are records of activities or events in various forms and media by developing information and communication technology, created and received by various institutions and organizations in community, national, and state life [1]. Dynamic archives are a type of archive that is directly used in the activities of the archive creator and is stored for a certain period. The management of incoming mail as part of dynamic archives is carried out to ensure the availability of relevant archives to support activities as materials for accountability and legitimate evidence [2]. According to the National Archives of the Republic of Indonesia Regulation Number 6 of 2021, efforts to register

and digitize archives are necessary to support the Electronic-Based Government System [3].

As part of a government agency, the Borobudur Subdistrict was chosen as the research location due to the need for archive registration and digitization, particularly in the General Administration Subdivision. Approximately 10,000 incoming letters have not been digitized over the past five years. With this amount, manual registration and digitization would take significant time. This was proven through the trial of registering and digitizing several incoming letters manually. The trial results showed that only 20 letters could be processed in one day. Therefore, a system is needed to digitize all incoming letters in the Borobudur Subdistrict.

The system is designed using the Optical Character Recognition (OCR) Tesseract technique, a technology that can convert image-based documents into digital text [4] and reduce the risk of data loss [5]. The generated text is then processed using the Regular Expressions method, which matches patterns with the required text [6][7]. Furthermore, the TextRank method summarizes the letter's content, as this technique can identify the most relevant parts of a document [8]. The application of OCR, Regular Expressions, and TextRank techniques in the archiving system at Borobudur Subdistrict is expected to improve the efficiency of the incoming mail archive digitization process.

In the system design, the Waterfall method is used. The Waterfall method is commonly employed in developing new information systems, where the implementation is carried out systematically and progressively, ensuring high-quality output [9]. Compared with the Rapid Application Development (RAD) method, which allows for flexible adjustments in response to environmental changes or user needs, the Waterfall method is suitable for developing this archiving system because it already has clear specifications and does not require many adjustments [10]. The Waterfall method was also chosen because the archiving system has a clear objective [11]. On the contrary, the Prototype method is unsuitable as it is more appropriate for systems built based on specific demands and needs [12].

Various previous studies on archiving systems have been conducted. Among them is the study by Indriyani & Hasibuan, which utilized OCR for digitizing incoming and outgoing letters at the Batang Toru Subdistrict Office, Tapsel [13]. A study by Andreas Gosal & Dolf Rompas demonstrated that the application of OCR significantly accelerates the archiving process and reduces the risk of human error [14]. The research by Bintang, Ashshidiq, & Dzakwan explain that regular expressions can be used for pattern-based searches [7]. Additionally, a study by Zamzam and Muhammad Adib show that document summarization automation can assist in the summarization process as it involves computer calculations, compared to manual summarization [8].

Previous research has shown that the application of OCR techniques, Regular Expression methods, and TextRank methods has proven effective in increasing the efficiency of the archive digitization and registration process. However, no single study

has integrated all of these techniques to further improve and optimize the speed of archive digitization and registration. Moreover, effective solutions for registering and digitizing large volumes of archives, such as those found in the Borobudur Subdistrict, have not yet been examined.

By implementing OCR techniques, along with Regular Expressions and the TextRank method, this study offers a novel solution for digitizing and registering archives in a government agency. Unlike previous research, this system is designed to handle large archives (10,000 records) with significantly greater efficiency. Moreover, this study has a clear objective and follows a systematic process by applying the Waterfall method.

Method

The research method used in this study is the Waterfall Method, which has stages shown in Figure 1. These stages include requirements Analysis, System Design, System Implementation, and Testing [15].

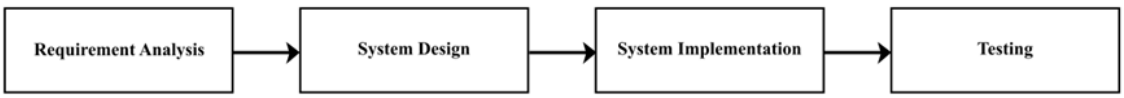

Figure 1. Waterfall Method Stages

Requirement Analysis

The requirements were identified through the involvement of relevant stakeholders, including archive officers and the office's leadership team. Data were gathered through in-depth interviews with stakeholders and the distribution of questionnaires to users. In addition, observations were made of the ongoing processes, and the necessary specification documents were developed. Following the completion of this stage, the following findings were obtained:

1. The system must include features for archive management, such as adding, deleting, editing, and viewing archives.
2. The system must be capable of reading and converting physical documents into digital text using Optical Character Recognition (OCR) technology.
3. The system must be able to detect specific patterns in the OCR extraction results using the Regular Expressions method.
4. The system must automatically generate summaries from the OCR extraction results using the TextRank method.

System Design

1. Use Case Diagram and Data Flow Diagram

Figure 2 presents the Use Case Diagram created to visually represent user and system interactions. Meanwhile, Figure 3 illustrates the Data Flow Diagram (DFD), a graphical

representation of the system that depicts its components, the flow of data between them, and the sources, destinations, and storage of that data.

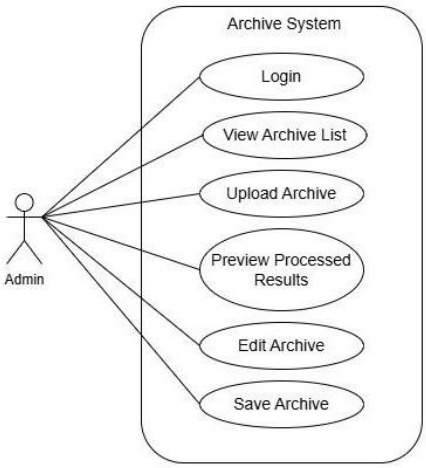

Figure 2. Use Case Diagram

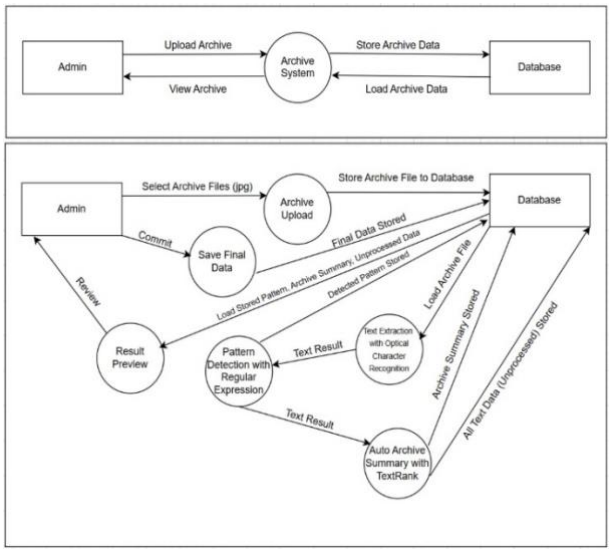

Figure 3. Data Flow Diagram (DFD)

2. Class Diagram

Figure 4 depicts the system's structural design. It presents the Class Diagram created to illustrate the system's structure and simplify the programming process. This diagram provides a detailed representation of the system's classes, including their attributes and relationships, offering a clear understanding of the system's architecture and its components' interactions.

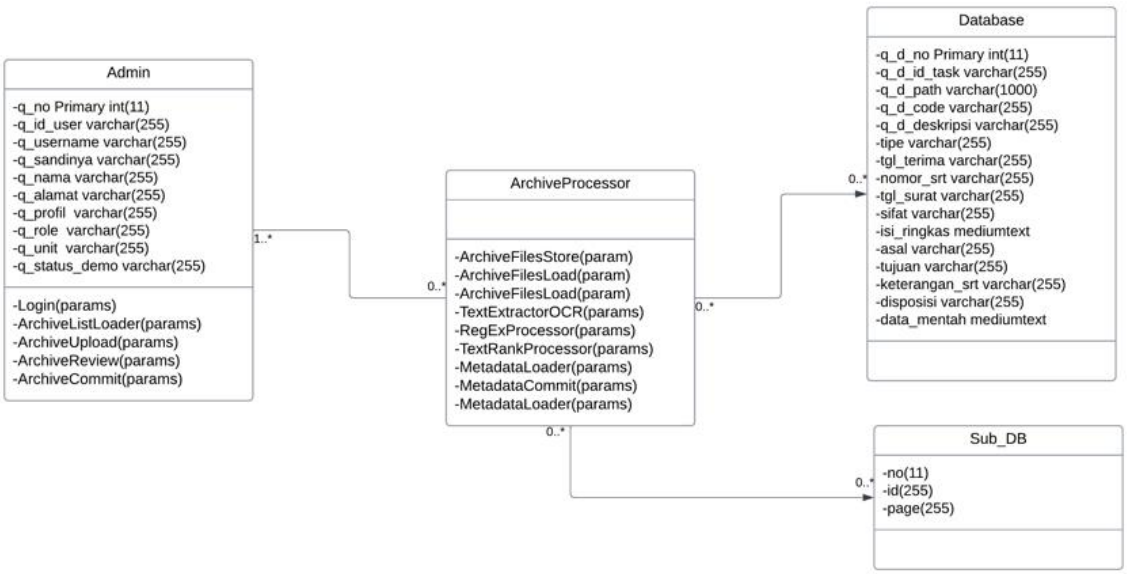

Figure 4. Class diagram

3. Activity Diagram

The Activity Diagram, shown in Figure 5, was designed to depict the workflow and steps of the system's processes from start to finish.

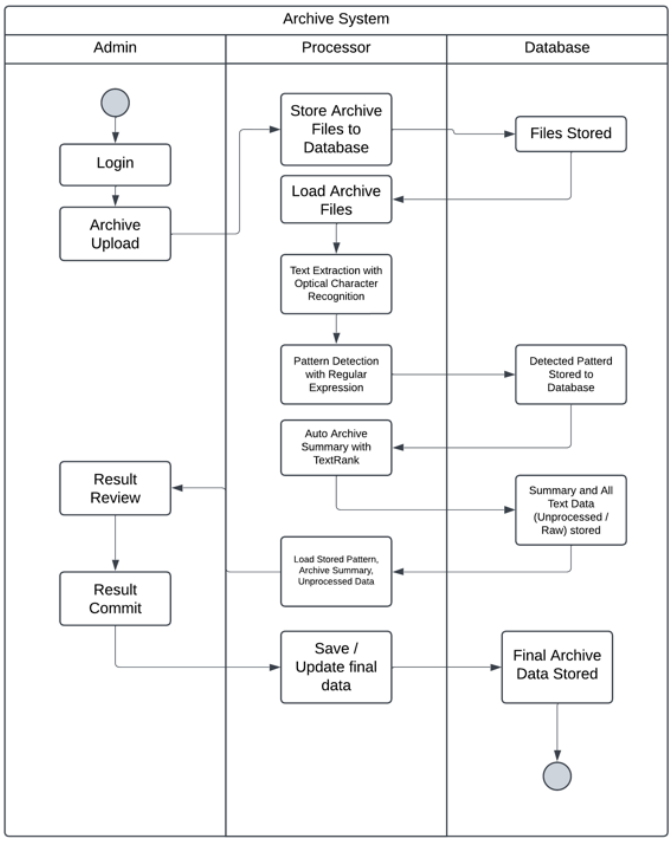

Figure 5. Activity diagram

System Implementation

1. Archive Upload

The first feature is document upload, as illustrated in Figure 6. The uploaded documents can be photos, and there can be more than one photo. The upload process must follow the page order.

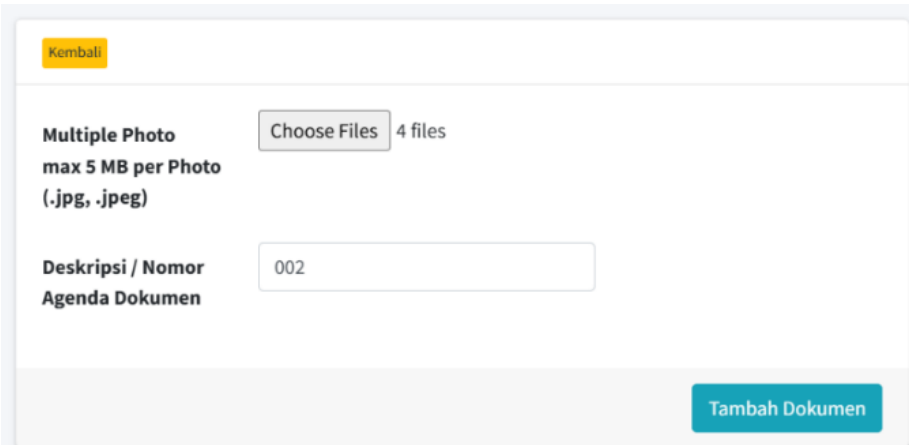

Figure 6. Archive upload interface

2. Optical Character Recognition (OCR) Result

From the uploaded archive, as shown in Figure 7, the system will extract the text to store it in the database.

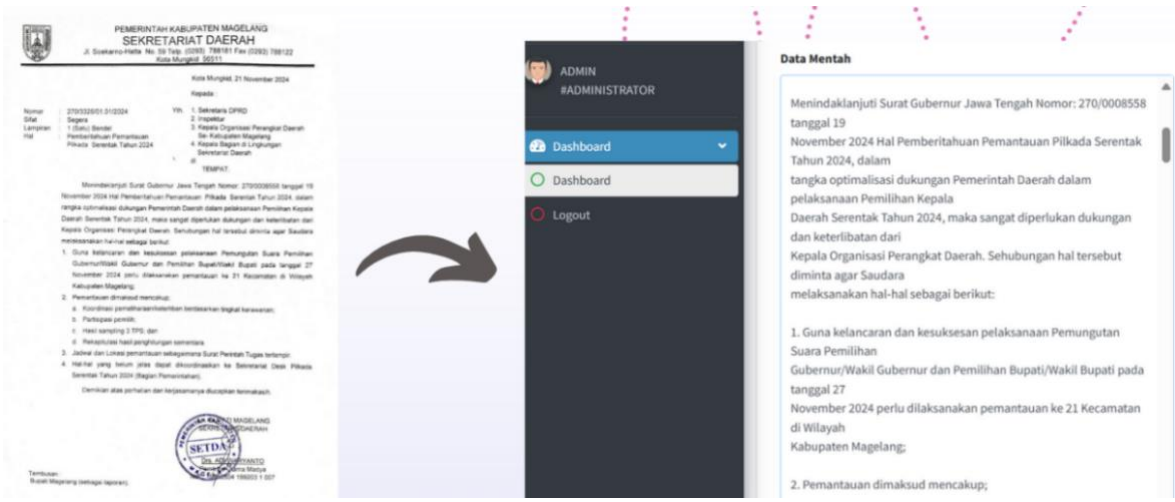

Figure 7. Optical Character Recognition (OCR) Result

3. Regular Expression Result

The OCR extraction results are then processed using the Regular Expressions method, as illustrated in Figure 8, allowing the desired pattern to be identified.

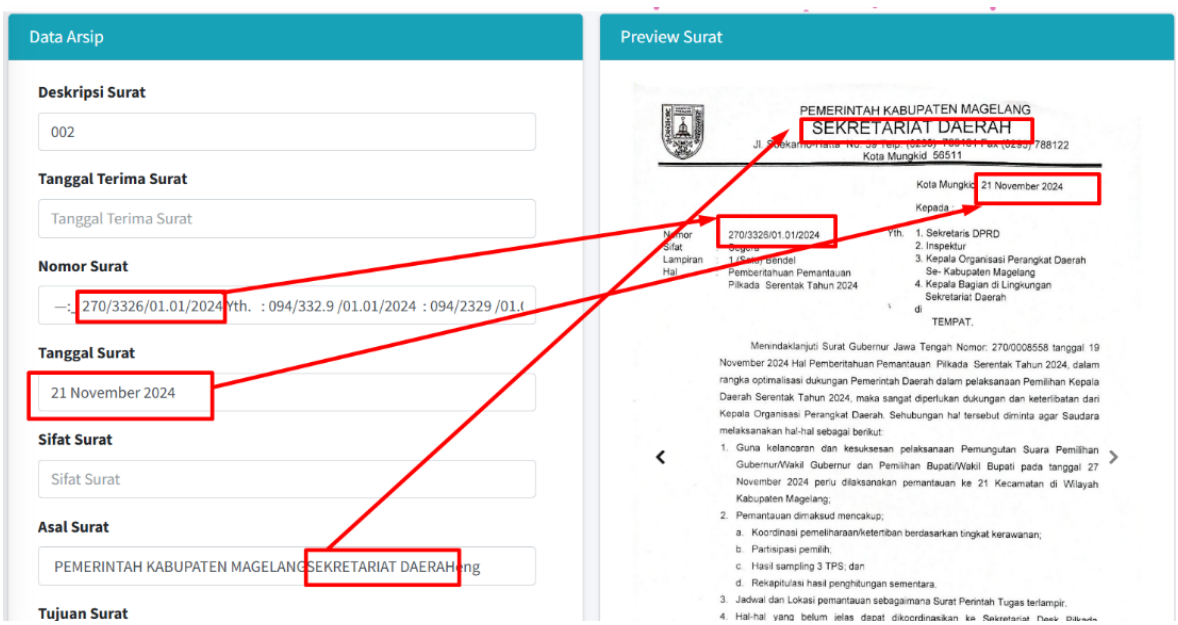

Figure 8. Regular Expression Method Result

4. Editing Section

The patterns identified through the Regular Expression process can be reviewed and updated as needed via the editing page, as depicted in Figure 9. This provides flexibility for making adjustments to the extracted data.

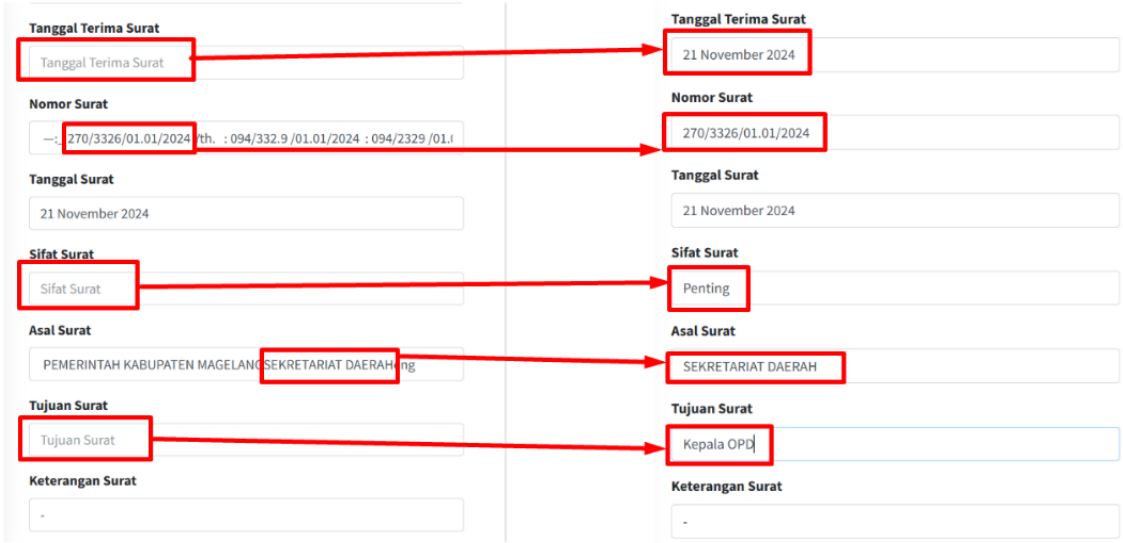

Figure 9. Edition section

5. TextRank Result and Final Digital Archive Result

Finally, as presented in Figure 10, the archive data will be entered into the database as the final data. A summary of the archive content has also been automatically generated with TextRank Methods.

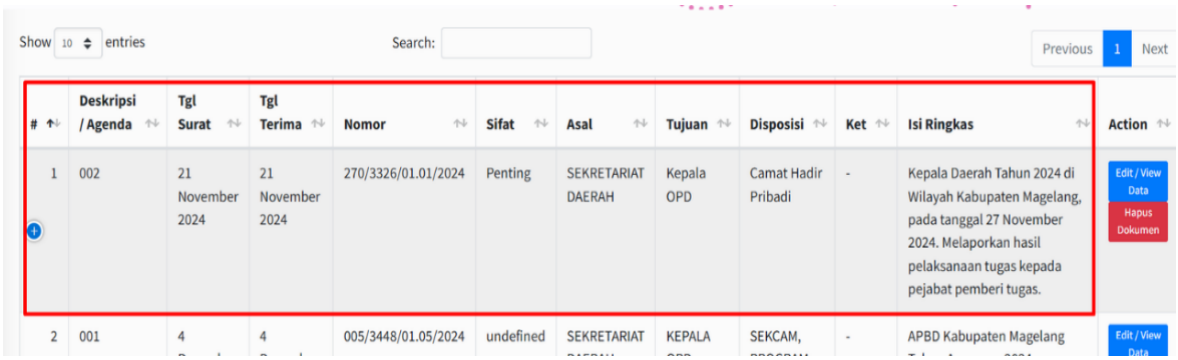

Figure 10. TextRank Result and Final Digital Archive Result

Testing

The testing was conducted by having archive officers use the developed system. The officers completed the digitization process with the same time allocation as during conventional digitization. On average, officers allocated 2 hours per day for digitization tasks. After using the system, the officers could digitize 80 archives within approximately 2 hours.

Result and Discussion

System Implementation Result

The implementation of the archive digitization system using OCR, Regular Expression, and TextRank methods shows satisfactory results in terms of processing speed. The test results indicate that the system can process an average of 80 letters per day, a

significant increase compared to the manual method, which could only process 20 letters per day.

Discussion

Calculating the efficiency increase can be done by comparing the results before and after implementing the system, which helps measure the improvement in processing capacity.

The formula for calculating the efficiency increase is shown in Figure 11.

$$\text{Percentage Increase} = \left(\frac{\text{New Output} - \text{Old Output}}{\text{Old Output}}\right) \times 100$$

Where:

- **Old Output** = 20 archives per day
- **New Output** = 80 archives per day

$$\text{Efficiency Increase} = \left(\frac{80 - 20}{20}\right) \times 100 = 300\%$$

Figure 11. Calculating the Efficiency

The results show a significant increase in efficiency, up to 300%, where the number of archives registered and digitized in one day increased from 20 to 80.

These findings demonstrate that OCR techniques can assist and accelerate the archiving process by performing text extraction. Then, it has also been proven that archive processing and summarization become much more effective and efficient when the Regular Expression method for pattern matching is combined with the TextRank method for automatic summarization. Furthermore, the Waterfall method used in the system design ensures a systematic development process and is emphasized for its suitability in projects with well-defined requirements and objectives.

The system created is developed by combining all of them, including OCR techniques, Regular Expressions, and TextRank methods, and through a structured development process. This system can effectively and efficiently address the challenges of digitizing and registering large archives in the Borobudur Subdistrict. These results align with the primary objectives of this research.

Conclusion

Implementing the archive digitization system utilizing OCR, Regular Expression, and TextRank methods has significantly enhanced the efficiency of archive processing in the Borobudur Subdistrict. With an observed increase in processing speed from 20 to 80 letters per day, the system demonstrates a 300% improvement in efficiency. This result demonstrates the system's capability to provide an effective solution for the registration and digitization process. Further research can be conducted to improve efficiency in the digitization process by enhancing the capability and accuracy of OCR,

Regular Expression, and TextRank. This could bypass manual review or editing after document upload, making the process fully automated

References

[1] Arsip Nasional Republik Indonesia Peraturan Arsip Nasional Republik Indonesia Nomor 4 Tahun 2021 Tentang Pedoman Penerapan Sistem Informasi Kearsipan Dinamis Terintegrasi; Indonesia, 2021; pp. 1–18;.

[2] Arsip Nasional Republik Indonesia Arsip Dinamis. Arsip Nasional Republik Indonesia Peraturan Arsip Nasional Republik Indonesia Nomor 6 Tahun 2021 Tentang Pengelolaan Arsip Elektronik; 2021; Vol. 1, pp. 1–24;.

[3] Kusmanto, B.T.; Pradana, N.; Prakisya, T.; Hatta, P. Comparative Analysis of Google Vision OCR with Tesseract on Newspaper Text Recognition. Media Comput. Sci. 2024, 1, 31–46, doi:10.69616/mcs.

[4] Dermawan, M.S.; Mulyawan, B.; Lauro, M.D. Perancangan Aplikasi Sistem Manajemen Dokumen Dan Pencarian Teks Dengan Menggunakan Optical Character Recognition (OCR). J. Ilmu Komput. dan Sist. Inf. 2019, 7, 81–86.

[5] Chen, Q.; Banerjee, A.; Demiralp, Ç.; Durrett, G.; Dillig, I. Data Extraction via Semantic Regular Expression Synthesis. Proc. ACM Program. Lang. 2023, 7, doi:10.1145/3622863.

[6] Bintang, J.M.; Ashshidiq, M.F.; Dzakwan, H.F. Penerapan Algoritma String Matching Dan Regular Expression Pada Aplikasi Kamus Besar Bahasa Indonesia (KBBI). BIOS J. Teknol. Inf. dan Rekayasa Komput. 2023, 4, 34–41, doi:10.37148/bios.v4i1.57.

[7] Zamzam, M.A. Sistem Automatic Text Summarization Menggunakan Algoritma Textrank. Matics 2020, 12, 111–116, doi:10.18860/mat.v12i2.8372.

[8] A. A. Wahid "Analisis Metode Waterfall Untuk Pengembangan Sistem Informasi," . J. Ilmu-ilmu Inform. dan Manaj. STMIK 2020, 1.

[9] Shandra Dewi, E.; Ardya Mesia Putri, E.; Tji Beng, J.; Teknologi Informasi, F. Perbandingan Antara Metode Waterfall Dan Metode Rad Dalam Pembuatan Aplikasi E-Rekrutmen Berbasis Website: Studi Kasus Pt Xyz Comparison Between the Waterfall Method and the Rad Method in Creating Website-Based E-Recruitment Applications: A Case Study Of . J. Inf. Technol. Comput. Sci. 2024, 7, 1067–1072.

[10] Murdiani, D.; Sobirin, M. Perbandingan Metodologi Waterfall Dan Rad (Rapid Application Development) Dalam Pengembangan Sistem Informasi. JUTEKIN (Jurnal Tek. Inform. 2022, 10, doi:10.51530/jutekin.v10i2.655.

[11] Widya Ningsih PERBANDINGAN MODEL WATERFALL DAN METODE PROTOTYPE UNTUK PENGEMBANGAN APLIKASI PADA SISTEM INFORMASI. J. Ilm. Metadata, 2023, 1, 83–95, doi:10.62386/jised.v2i1.50.

[12] Hasibuan, H. novita sari Optical Character Recognition Untuk Manajemen Surat. (CoSIE) 2022, 01, 146–151.

[13] Andreas Gosal, F.; Tinno Dolf Rompas, P. Penerapan Teknologi Optical Character Recognition Pada Pengarsipan Dokumen (Studi Kasus: PT Pertamina Geothermal Energy Area Lahendong). Innov. J. Soc. Sci. Res. 2023, 3, 5404–5422.

[14] Arief, M.; Budi, S.; Sadiah, H.T. Digitalisasi Pengarsipan Surat Pada Kantor Kecamatan Cigudeg Digitalizing Letters in the Kecamatan Office of Cigudeg. Bisnis dan Komput. 2021, 1, 38–43.