Pak. j. life soc. Sci. (2025), 23(2): 1173-1182 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences

www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2025-23.2.00245

RESEARCH ARTICLE

Intelligent Document Recognition System in The Context of Digitalization of Social Work

Natalia Yazvinskaya[1]*, Svetlana Belikova[2]

[1, 2]Don State Technical University, Rostov-on-Don, Russia

ARTICLE INFO

Received: Oct 14, 2025

Accepted: Dec 9, 2025

Keywords

Digital Inequality
Document Management
Software Automation
Algorithm

***Corresponding Author:**

lionnat@mail.ru

ABSTRACT

Despite technical readiness, the process of full digitalization of document management faces a number of fundamental challenges. This paper analyzes four key groups of problems: regulatory and legal (direct legislative restrictions, the problem of long-term authenticity of electronic documents and interdepartmental interaction), technological (cybersecurity, system compatibility, digital inequality), organizational and economic (high implementation costs, business process transformation) and socio-cultural (psychological resistance, low digital literacy). Based on the conducted research, conclusions were drawn about the need for a synchronized development of the legal framework, technological infrastructure and educational programs, and a practical implementation of software for various stakeholders was proposed. The presented software is aimed at automating the electronic registration of copies of incoming documents and will allow social organizations to significantly accelerate the transition to digital document management. A software algorithm has been developed that reduces application runtime and decreases system resource consumption. Three interfaces are proposed: authorization, user interface and administrator interface. The software functionality is demonstrated and recommendations for its possible application are given.

INTRODUCTION

The modern world would be impossible without the widespread use of information technology in all spheres of life. One of the first areas of automation sought in the early stages of the digital revolution was the automation of document management systems. The implementation of modern document management systems in various fields, while adhering to recognized standards for the integration of electronic document management, becomes a powerful tool for organizational development, significantly reducing the time and cost of document processing and increasing the speed and efficiency of work overall.

However, the idea of completely eliminating paper from document management systems is still largely utopian at this stage of development. This is not a technical impossibility, but a complex systemic shift affecting technological, legal, human, and cultural aspects.

Regulatory barriers: in the grip of outdated paradigms

Despite the formal recognition of the legal force of electronic documents and electronic signatures, current legislation in many countries, including the Russian Federation, still contains direct and indirect references to the mandatory nature of paper documents, creating legal conflicts and systemic mistrust of the digital environment.

Direct legislative restrictions (“paper mandate”)

Current archival storage regulations are primarily focused on paper documents. Existing rules and standards govern the description, accounting, and preservation of traditional paper media in much greater detail, leaving electronic document management issues underdeveloped. For example, international and Russian archival standards permit the storage of documents in both paper and

electronic form, but the choice depends on the type of document and its origin [1-3]. Documents created electronically, such as those signed with an Electronic Digital Signature (EDS), must be stored electronically using specialized storage systems and preserving the Electronic Signature (ES) to confirm their authenticity and integrity. Paper documents are stored in accordance with archival regulations (temperature, humidity, protection from light), while electronic documents created on paper can be scanned and converted to electronic format for easy storage, while preserving the originals or their copies.

This creates significant challenges for organizations seeking to fully digitalize their document management systems. Fundamental documents on archival management focus primarily on working with paper media, which influences approaches to developing document storage and accounting systems [1-4]. At the same time, electronic document management issues require additional development and standardization to ensure their effective implementation in organizational practices.

This situation creates a certain imbalance between modern digitalization trends and current regulatory requirements, complicating the transition of organizations to fully electronic document management systems.

Similar problems arise when submitting documents to court [5]. Courts are not always willing to accept electronic versions of documents and often insist on the submission of paper copies of documents used in legal proceedings.

The problem of long-term authenticity and legal force

An electronic signature is tied to Cryptographic Protection Tools (CPS) and Certification Authority (CA) certificates [6-7]. Encryption algorithms quickly become obsolete as hardware performance increases. Regulators require a timely transition to new modern digital signature and hashing standards, which raises the issue of technological obsolescence. Suppose the authenticity of a document signed with an EDS needs to be verified 15-20 years from now. The question of how to verify its authenticity if the CA ceases to exist or the encryption algorithm is deemed insecure by that time remains unresolved. This problem does not exist for paper documents; if the paper on which the document is written has not decayed, then an examination of the signature and document's authenticity is possible centuries later.

The problem of interdepartmental and international cooperation

The lack of clear, internationally enshrined mechanisms for verifying the authenticity of an electronic document throughout its entire lifecycle creates risks for businesses and individuals [8].

Even within a single state, different agencies may have technically different requirements for file formats, types of digital signatures, and exchange protocols. This forces organizations to interact with each agency separately, minimizing the effectiveness of paperless communication.

Within the Eurasian Economic Union, a mechanism for recognizing digital signatures in electronic documents was developed to ensure the legal validity of electronic documents during cross-border information exchange. This mechanism is based on the task of confirming the authenticity of electronic documents and digital signatures of legal entities, accomplished using a trusted third-party service [6]. A similar "island" of mutual recognition of digital signatures also exists in the European Union under the eIDAS Directive [5]. However, the creation of a unified legal framework for mutual recognition of digital signatures is not yet underway due to various technological and political reasons. Thus, in most countries, the legal validity of the Russian Qualified Electronic Signature (QES) is not recognized. To conduct international economic activity, companies are forced to duplicate their document flow on paper to ensure its legal validity worldwide. The review and analysis presented key barriers to the transition to paperless document flow (Figure 1).

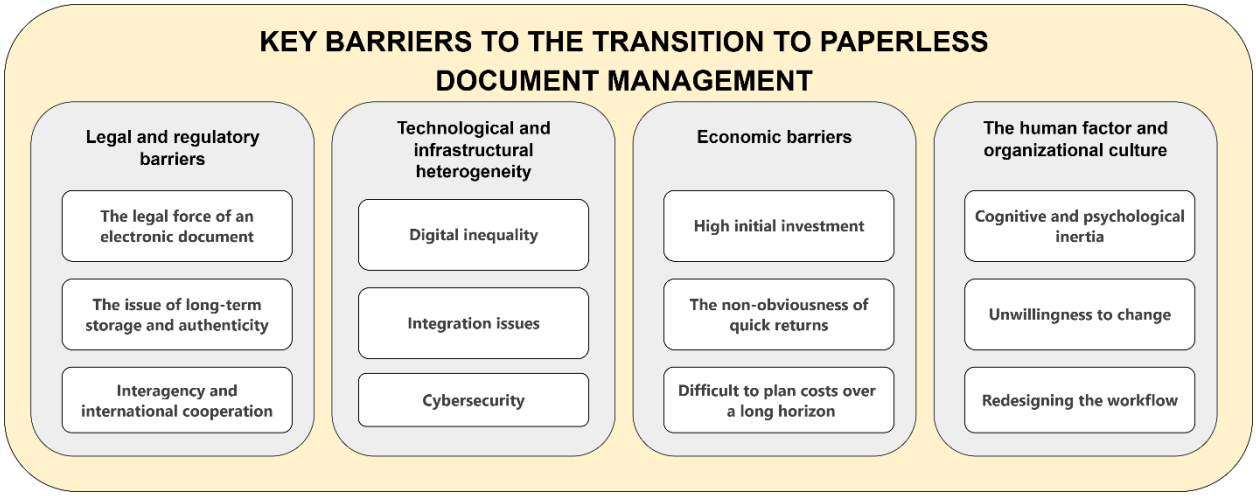

Figure 1: Barriers to the transition to paperless document management

Legal barriers are systemically important. They don't simply slow down the transition; they legally perpetuate the need for paper in key areas. What's needed is not just targeted legislative amendments, but a comprehensive review of all legislation to address the "paper mandate" and the adoption of new, proactive legislation in the area of digital law, including regulation of trusted technologies and the long-term storage and verification of digital signatures. Without addressing these issues at both the national and international levels, any technological initiatives will be hampered by legislative imperfections.

Technological and infrastructural heterogeneity

Technological readiness for paperless document management is extremely uneven both across and within different economic sectors. This heterogeneity creates a "digital divide," where advanced organizations are already reaping the benefits of digitalization, while others remain trapped in the paper era due to high entry costs, integration difficulties, and unresolved fundamental issues related to data security and integrity.

Digital divide

The corporate sector, small and medium-sized businesses, and the public sector have vastly different resources and capabilities for transitioning to paperless document management. Large corporations have the financial and human resources to implement expensive planning and electronic document management systems, deploy their own or lease secure data centers, and maintain a staff of IT specialists and information security administrators. However, for small businesses, individual entrepreneurs, and many public institutions (schools, clinics, municipal administrations), these costs are prohibitive. They are forced to either continue working in the old way or use insecure, makeshift methods. Exchanging scanned documents via email and sending copies of documents via various messaging apps is still widely used.

The quality and speed of internet connections, as well as access to IT services, can vary dramatically across regions. This creates a barrier for organizations in small towns and rural areas, where full technical capabilities for electronic collaboration are lacking.

The problem of integration and compatibility

The lack of unified, mandatory national standards for data formats (XML, JSON, ODF, PDF) and exchange protocols leads to a situation where different systems cannot "understand" each other [8-9]. One organization's accounting software cannot directly retrieve data from another organization's customer relationship management system. This requires the development of expensive custom connectors or manual data transfer, which negates the effectiveness of digital information exchange. The costs of integrating disparate systems often exceed the cost of acquiring the software products themselves many times over. For an organization using 10-15 different programs, creating a unified information space becomes an overwhelming project, requiring prohibitive implementation costs.

Cybersecurity – Vulnerability as a systemic risk

In a paper-based document management system, the compromise of a single document is a rare occurrence, easily handled by any security service. In a centralized digital system, a successful cyberattack can paralyze a company's entire document management system, bringing all operations to a complete halt. Furthermore, the costs of mitigating the consequences can have a fatal impact on the company's continued existence [10].

Ensuring document authenticity throughout its lifecycle and confirming its integrity are critically complex. This requires the implementation of complex and expensive blockchain-based solutions or trusted third-party services.

In a paperless document management system, the risks of social engineering, employee negligence (such as the use of weak passwords, susceptibility to phishing attacks, and disregard for information security regulations) remain, as they do in a mixed document management system. However, the significance of their consequences increases dramatically in a digital environment.

The problem of long-term storage and technological obsolescence

File formats, software, and storage media become obsolete within 5-15 years. Ensuring that a document in .docx format or signed with an outdated digital signature certificate can be securely opened and authenticated 25 years later is a complex and expensive task, requiring constant data migration to new media and legitimate conversion to new formats, which generates additional operational costs and the risk of errors during the transfer.

Current office management and archival regulations are primarily focused on the storage of paper originals [8]. Legal procedures for the creation and transfer of electronic files for permanent storage in state archives are insufficiently detailed and require the transfer of containers with electronic documents to the state archive on physically separate physical media, such as hard drives, which requires additional material costs and reduces the effectiveness of the transition to paperless document management. Technological barriers are the most obvious, but overcoming them is hampered not so much by the lack of solutions as by their cost, complexity, and the immaturity of the regulatory framework. Infrastructure heterogeneity creates an environment were finding a universal solution for everyone is quite difficult. This requires the government to implement programs to support small and medium-sized businesses and the public sector, develop uniform standards and protocols, and create a national strategy for long-term digital data preservation.

Development of a universal document processing module in various formats for the implementation of paperless document management

Taking into account the above-described issues, a universal algorithm for processing documents in various formats was developed, ensuring acceptable response times and low hardware platform requirements for deployment (Figure 2). A properly designed algorithm simplifies software development and accelerates application debugging. Well-organized program code, in turn, simplifies adjustments and support.

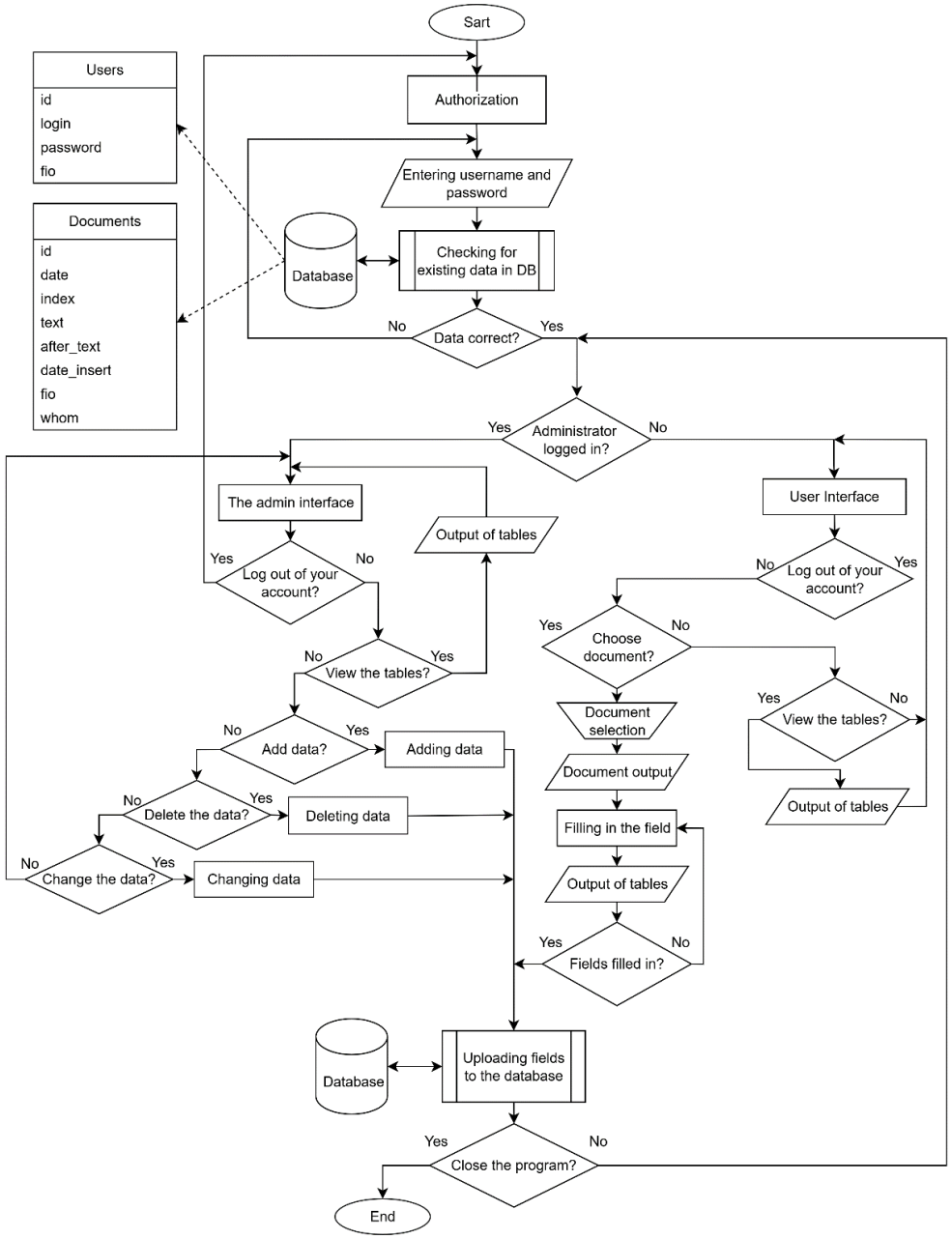

Figure 2: Application operation algorithm

METHODS AND TECHNOLOGIES

The primary criteria for selecting the technology stack were functionality, usability, stability, and software support. Each of the tools was chosen based on its features and development approach: C#, PostgreSQL, Visual Studio, DataGrip, and the Tesseract library.

Implementation of the algorithm

The application has three interfaces: an authorization interface, a user interface, and an administrator interface.

After launching the application, authorization begins: the user enters their login and password, which are verified in the database. Depending on the user's role, the interface displayed is: the administrator can view and edit tables; the user can register documents and view tables. The administrator can edit tables, delete data older than a specified age, and manage users and data in the "users" table (MD5 passwords). After logging out, the user returns to the authorization window; the administrator interface is illustrated in Figure 3.

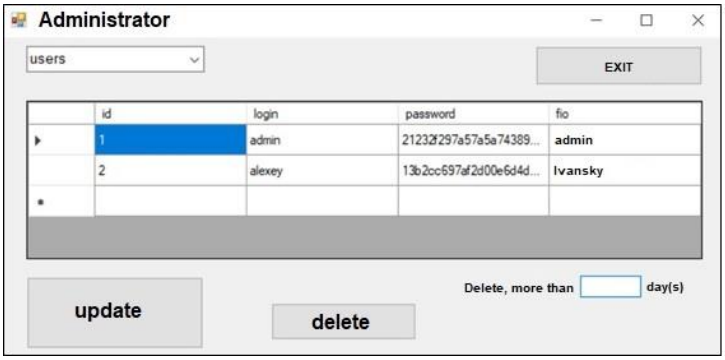

Figure 3: Administrator interface

The user interface is limited to viewing tables. The primary function is document registration: select a document, its contents, and automatically populate the type, date, index, and recipient fields for the memo. The user completes the content and notes fields. If the automatic fields are incorrectly formatted, a warning appears. After filling in the data, use the "Add Data" button to add it. The table can be updated for re-upload, and the tables are automatically populated with the last name, first name, patronymic name of the document controller, and the registration date. The user interface is shown in Figure 4.

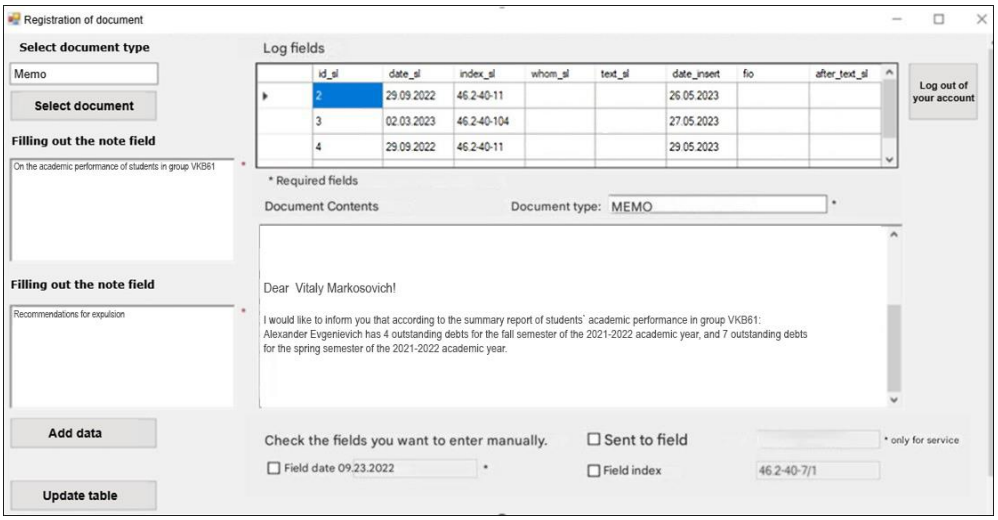

Figure 4: User interface

Figure 5 shows an option for storing information about documents in a database.

	id_sl	date_sl	index_sl	whom_sl	text_sl	after_text_sl	date_insert
1	2	29.09.2022	46.2-40-11				26.05.2023
2	3	02.03.2023	46.2-40-104				27.05.2023
3	4	29.09.2022	46.2-40-11				29.05.2023

Figure 5: Database table

The developed program has the ability to keep track of documents not only in .doc and .docx formats, but also scanned images of documents in .pdf format and screenshots of documents or their photographs in .png or .jpg formats.

RESULTS

Additionally, when uploading images or PDF files, the user has the option to view their contents not only as the uploaded text in the document content field, but also as an image. Recognition does not provide a complete guarantee that all document elements are recognized correctly by the Tesseract library.

Scanned documents and images may contain handwritten text. While the Tesseract library is good at recognizing printed Latin characters, it is significantly more difficult to use the library for recognizing handwritten text [13]. This is why the manual document field entry feature was

introduced in the user interface. An example of an incoming document page is shown in Figure 6. The output of the scanned document contents with handwritten text is shown in Figure 7.

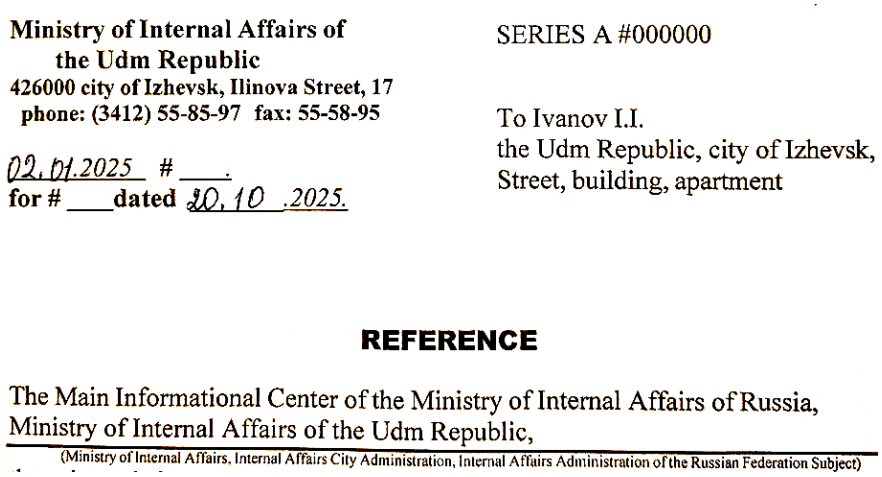

Figure 6: Scanned image of the document

Figure 7: Output of the scanned document

This example confirms the presence of problems in recognizing handwritten characters, while the rest of the printed text was recognized completely correctly.

PDF documents can be divided into searchable and non-searchable. Searchable PDF documents contain text content that can be processed by search engines. They can be created from native text files, such as Microsoft Word, HTML, or full-text search platforms.

For easy searching and use of such documents, it is recommended to use PDF files with extended functionality, such as PDF/A and PDF/X, which support text searching, indexing, and archiving. Furthermore, such documents are easily processed by automation programs.

Non-searchable PDF documents are typically created from image files such as JPG or PNG. They do not contain text information, but only images or page scans. Since search engines cannot process image files as text information, such PDF files cannot be processed using search tools.

The developed program takes into account the document selection condition: if it is a searchable PDF document, its information will be displayed as text in the document content field, without converting its pages to images.

If the PDF document is not searchable, its conversion to images will be handled by the dedicated Docotic.pdf library, while image recognition and text extraction are handled by the Tesseract library. Tesseract searches for patterns in pixels, letters, words, and sentences, using a two-step approach called habitual recognition. It performs one pass through the data to find characters, then a second to fill in any letters it's unsure about with letters that are more likely to match the corresponding pass or sentence context. The function for converting document pages to images and extracting text from them is shown in Figure 8.

```
{
    var StrBuildDoc = new StringBuilder();
    using (var pdf = new BitMiracle.Docotic.Pdf.PdfDocument(filepath))
    {
        using (var tesseract = new TesseractEngine(@"tessdata", "rus", EngineMode.LstmOnly))
        {
            tesseract.SetVariable("textord_min_linesize", 2.5);
            tesseract.SetVariable("lstm_choice_mode", 2);
            for (int i = 0; i < pdf.PageCount; ++i)
            {
                if (StrBuildDoc.Length > 0)
                    StrBuildDoc.Append("\r\n\r\n");

                BitMiracle.Docotic.Pdf.PdfPage page = pdf.Pages[i];
                string readingtext = page.GetText();
                if (!string.IsNullOrEmpty(readingtext.Trim()))
                {
                    StrBuildDoc.Append(readingtext);
                    continue;
                }

                foreach (BitMiracle.Docotic.Pdf.PdfImage image in page.GetImages())
                {
                    if (image.Height == 512)
                        image.ReplaceWith("1px.png");
                }

                PdfDrawOptions Drawoptions = PdfDrawOptions.Create();
                Drawoptions.BackgroundColor = new PdfRgbColor(255, 255, 255);
                Drawoptions.HorizontalResolution = 100;
                Drawoptions.VerticalResolution = 100;

                string pageSave = $"C:/Users/jigul/Desktop/Диплом/scaning_page/page_{i}.png";
                string pageImage = $"page_{i}.png";
                page.Save(pageImage, Drawoptions);
                page.Save(pageSave, Drawoptions);
                using (Pix img = Pix.LoadFromFile(pageImage))
                {
                    using (Tesseract.Page recognizedPage = tesseract.Process(img))
                    {
                        string recognizedText = recognizedPage.GetText();
                        StrBuildDoc.Append(recognizedText);
                    }
                }

                File.Delete(pageImage);
            }
        }
    }

    using (var writer = new StreamWriter("result.txt"))
    {
        writer.Write(StrBuildDoc.ToString());
        richTextBox3.Text = StrBuildDoc.ToString();
```

Figure 8: Function for extracting text from a PDF document

Otherwise, the program will extract text from .doc and .docx documents seamlessly, line by line, just as it was in the document itself.

Regular expressions are used for automatically populating and validating text fields. The program includes regular expressions for matching document names and several date formats.

Before automatically populating fields, the program checks each document line against the regular expression, which may contain the date and document index. If such a line is found, the line is checked again for the correct date format, and after a successful check, all data is entered into the fields. The data validation function for .doc and .docx documents is shown in Figure 9.

```
if (Regex.IsMatch(textBox4.Text, prikaz))
{
    string regul_all_prik = @"^\D[0-9]{2}\D\s[а-я]+\s[0-9]{4}\s[г]?[.]?\s+[№]\s\S+\s*$";
    string regul_date_prik = @"^\D[0-9]{2}\D\s[а-я]+\s[0-9]{4}$";
    int k = 0;

    while (!Regex.IsMatch(richTextBox3.Lines[k], regul_all_prik, RegexOptions.IgnoreCase))
    {
        k++;
        if (Regex.IsMatch(richTextBox3.Lines[k], regul_all_prik, RegexOptions.IgnoreCase))
        {
            break;
        }
    }
    string dateDoc_else_line = richTextBox3.Lines[k];
    if (Regex.IsMatch(dateDoc_else_line, regul_all_prik, RegexOptions.IgnoreCase))
    {
        string[] dateDoc_else1 = dateDoc_else_line.Split(' ');
        string dateDoc_elseall = dateDoc_else1[0] + " " + dateDoc_else1[1] + " " + dateDoc_else1[2];
        if (Regex.IsMatch(dateDoc_elseall, regul_date_prik, RegexOptions.IgnoreCase))
        {
            textBox1.Text = dateDoc_elseall;
            textBox2.Text = dateDoc_else_line.Split().Last();
        }
        else
        {
            MessageBox.Show("Поля не соответствуют формату, пожалуйста, введите вручную.");
        }
    }
    else
    {
        MessageBox.Show("Поля не соответствуют формату, пожалуйста, введите вручную.");
    }
```

Figure 9: Data validation function

To successfully register a document in the database, it is necessary to correctly construct a query for populating fields.

To do this, the values filled in the text fields were written to variables, which were then used in the query for populating table fields.

The administrator has a slightly different form for filling, editing, and deleting data in tables.

Special events were created for the dataGridView1 table display form, which, when triggered, triggered special functions. For example, if the administrator clicked on a cell and all other cells in that row were empty, the function for creating a new row is executed. If at least one cell in that row is already filled, the function for editing is executed, without adding a new row. The delete function is triggered when the administrator selects the entire row and deletes it.

CONCLUSION

The transition to paperless document management is an evolutionary, not revolutionary, process. It requires both targeted measures to develop specific libraries, universal modules, and applications, as well as coordinated efforts at all levels: from harmonizing legislation and developing international standards to investing in infrastructure and extensive educational efforts.

Registering electronic documents allows for their systematization and makes searching and managing them more convenient and efficient. Therefore, the software developed for automating the electronic registration of copies of incoming documents will enable organizations to significantly accelerate the transition to digital document management.

REFERENCES

1.ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd Edition https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd (accessed on 27 August 2025).
2.Describing Archives: A Content Standard (DACS) https://mysaa.archivists.org/productdetails?id=a1B5a00000heUDGEA2 (accessed on 27 August 2025).
3.Rules for organizing the storage, acquisition, accounting and use of documents of the Archival Fund of the Russian Federation and other archival documents in state and municipal archives, museums and libraries, scientific organizations https://archives.gov.ru/documents/rules/pravila-2020.shtml
4.Haider, A.; Aryati, B.; Mahadi, B. "Opportunities and Challenges in Implementing Electronic Document Management Systems". Asian J. Appl. Sci. 2021, 3, 36–39. Available online: https://www.ajouronline.com/index.php/AJAS/article/view/2239 (accessed on 27 August 2023).
5.Mirosław Kutyłowski, Przemysław Błaśkiewicz, Advanced Electronic Signatures and eIDAS – Analysis of the Concept, Computer Standards & Interfaces, Volume 83, 2023,103644,ISSN 0920-5489, https://doi.org/10.1016/j.csi.2022.103644.(https://www.sciencedirect.com/science/article/pii/S0920548922000216)
6.6.Decision of the Board of the Eurasian Economic Commission of August 22, 2023 No. 120 "On the Rules for recognizing an electronic digital signature (electronic signature) in an electronic document and ensuring the legal force of electronic documents in cross-border information interaction of legal entities (business entities) with authorized bodies of the member states of the Eurasian Economic Union and the Eurasian Economic Commission using a trusted third party service" (https://www.consultant.ru/document/cons_doc_LAW_455537/92d3e3d03094ed76da5c15fa72b687f1cebd5931/).
7.Gokhool, O.; Nagowah, S.D. A Requirement Gathering Framework for Electronic Document Management Systems. In Proceedings of the IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 8–10 December 2022; pp. 1–6. Available online: https://ieeexplore.ieee.org/document/10037540 [Google Scholar] [CrossRef]

8.Sternad Zabukovšek S, Jordan S, Bobek S. Managing Document Management Systems' Life Cycle in Relation to an Organization's Maturity for Digital Transformation. Sustainability. 2023; 15(21):15212. https://doi.org/10.3390/su152115212

9.Rolland, K.H.; Hanseth, O. Managing Path Dependency in Digital Transformation Processes: A Longitudinal Case study of an Enterprise Document Management Platform. Sci. Direct 2021, 181, 765–774. Available online: https://www.sciencedirect.com/science/article/pii/S1877050921002726 (accessed on 15 May 2023).

10.Digital terror epidemic: two major companies have already collapsed from hacker attacks in a week (https://www.securitylab.ru/news/562021.php)

11.Mushhad, S., Gilani, M., Ahmed, J., & Abbas, M. A. (2009). Electronic document management: A paperless university model. In Proceedings – 2009 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT) (pp. 440-444). Article 5234679 IEEE. https://doi.org/10.1109/ICCSIT.2009.5234679

12.G. Abdul Robby, Antonia Tandra, Imelda Susanto, Jeklin Harefa, Andry Chowanda,Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application,Procedia Computer Science,Volume 157,2019,Pages 499-505,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.09.006. (https://www.sciencedirect.com/science/article/pii/S1877050919311640)

13. Tesseract documentation. URL: https://tesseract-ocr.github.io/