

Technological development of functionalities with convolutional neural networks for intelligent document management

Desarrollo tecnológico de funcionalidades con redes neuronales convolucionales para una gestión documental inteligente

Erika Marcela Parra-Amaya¹
Saida Ivonne Rojas-González²
Iliana Quintero-Percy³
Claudia Cristina González-Béndiksen De Zaldívar⁴
Nelson Giovanni Agudelo-Cristancho⁵

¹IMG (Colombia). Correo electrónico: eparra@imglatam.com
orcid: <https://orcid.org/0009-0009-6159-4491>

²IMG (Colombia). Correo electrónico: srojas@imglatam.com
orcid: <https://orcid.org/0009-0003-3834-9400>

³IMG (Colombia). Correo electrónico: iquinterop@imglatam.com
orcid: <https://orcid.org/0009-0009-1663-9670>

⁴IMG (Colombia). Correo electrónico: cgonzalez@imglatam.com
orcid: <https://orcid.org/0009-0006-9286-1757>

⁵Servicio Nacional de Aprendizaje SENA (Colombia).
Correo electrónico: nagudeloc@sena.edu.co
orcid: <https://orcid.org/0000-0002-1247-7696>

Recibido: 19-12-2024 Aceptado: 25-05-2024

Cómo citar: Parra-Amaya, Marcela; Rojas-González, Saida; Quintero-Percy, Ilian; González-Béndiksen De Zaldívar, Claudia; Agudelo-Cristancho, Nelson (2024). Technological development of functionalities with convolutional neural networks for intelligent document management. *Informador Técnico*, 88(1), 56-76. <https://doi.org/10.23850/22565035.6137>

Abstract

In the current highly digitized business landscape, efficient document management is crucial for increasing productivity and optimizing organizational processes. This need has been identified in this document, and an enhanced product named Infopoint, an enterprise content management software has been improved to provide intelligent document management through the implementation of advanced automation and artificial intelligence technologies. In document management, numerous processes are still manually executed, leading to errors and delays due to the burden of repetitive tasks. Manual management also causes internal delays and impacts interactions with clients, suppliers, and regulatory entities. Manual document searches consume valuable time and can result in unnecessary costs for the company. Infopoint addresses these challenges by incorporating automation features, particularly leveraging convolutional neural networks. This approach optimizes the functionality of incoming correspondence, reducing processing time by 29 %, on average. It also facilitates text and content searches within PDF documents, decreasing the average search time by 41 %. This article highlights how this improvement significantly reduces the time spent on correspondence management and information retrieval.

Keywords: convolutional neural networks; automation; machine learning; document management; enterprise content management software.

Resumen

En el contexto empresarial actual altamente digitalizado, la eficaz gestión de documentos es fundamental para aumentar la productividad y optimizar los procesos organizacionales. En este trabajo se ha identificado esta necesidad, y se ha potenciado un producto denominado Infopoint, un software de gestión de documentos electrónicos de archivo, que se ha mejorado para ofrecer una gestión documental inteligente mediante la implementación de tecnologías avanzadas de automatización e inteligencia artificial. En la gestión documental, numerosos procesos aún se llevan a cabo manualmente, lo que puede generar errores y demoras debido a la carga de tareas repetitivas. La gestión manual también ocasiona retrasos internos y en las interacciones con clientes, proveedores y entidades regulatorias. La búsqueda manual de documentos consume tiempo valioso y puede generar costos innecesarios para la empresa. Infopoint aborda estas problemáticas al incorporar funcionalidades de automatización, especialmente con el uso de redes neuronales convolucionales. Este enfoque optimiza la funcionalidad de correspondencia de entrada, reduciendo el tiempo de radicación en un 29 %, en promedio. También facilita la búsqueda de texto y contenido dentro de documentos PDF, disminuyendo el tiempo de búsqueda promedio en un 41 %. Este artículo destaca cómo esta mejora reduce significativamente el tiempo empleado en la gestión de la correspondencia y la búsqueda de información.

Palabras clave: redes neuronales convolucionales, automatización, aprendizaje automático, gestión documental, sistema de gestión de documentos electrónicos de archivo.

1. Introduction

Document management, as defined by the National General Archive (Archivo General de la Nación, 2013), involves a set of administrative and technical activities aimed at the planning, handling, and organization of documentation produced and received by entities, from its origin to its destination, to facilitate its use and preservation. To ensure control and security over an organization's entire document repository, including functions such as capture, management, storage, and distribution, electronic document management systems (EDMS) are employed (Rangel, 2017). One such system is the software Infopoint.

Within document management, numerous processes are often carried out manually. This not only introduces the possibility of errors due to the volume of repetitive tasks but also leads to dissatisfaction among employees responsible for these processes, as they are unable to dedicate their time to more valuable activities within the organization. A significant number of employees burdened with such tasks may find it challenging to unleash their creativity, resulting in decreased efficiency. According to a study conducted by Creative Live, an online platform offering diverse classes to boost creativity, approximately 30 % of employees in the United States would prefer a lower-paying job if it guaranteed more time for creative activities (Ricketts, 2014).

Manual handling of these processes can also cause delays and non-compliance with required procedures, internally and in dealings with clients, suppliers, and regulatory bodies. This, in turn, can lead to increased effort and time across different areas of the company to rectify errors and satisfy employees and/or clients who may be frustrated or inconvenienced by such situations (Llamas; López, 2020).

Similarly, employees within a company may spend considerable time manually searching, file by file, to locate a specific document. This search process can result in delays and additional costs for the company. A document management system may store thousands of documents of various types, such as text files, spreadsheets, images, etc. Searching or retrieving these documents beyond the basic metadata or keywords entered during classification, organization, and archiving can become challenging (Storecheck, 2022).

Thus, one of the issues with document management systems is the difficulty in conducting searches or queries within the stored documents, incurring costs in terms of time and resources that could be better utilized in

more productive tasks. This challenge may also contribute to non-compliance with internal and/or external requirements due to difficulties in finding specific documents, causing frustration for users engaged in time-consuming and operational document searches rather than focusing on more strategic tasks.

A study based on sources such as McKinsey & Company, the European Medicines Agency, and Forrester reveals that employees in companies that have not implemented technological solutions to enhance document control spend at least 20 % of their time searching for information. In other words, one out of every five employees dedicate their time to these delayed searches, therefore not contributing to the organization's productivity (Kantan Software, 2022).

As part of the continuous improvement process of Infopoint, there was an identified need to automate the most operational and manual tasks performed by software users, particularly regarding incoming correspondence. This optimization aimed to enable text and content searches within PDF documents, achieving intelligent document management. Technologies such as automation and long short-term memory (LSTM)¹ neural networks² were employed to enhance efficiency and effectiveness in document management, reduce costs, decrease task execution times, and minimize errors.

This article aims to demonstrate the use of convolutional³ neural networks to automate the functionality of incoming correspondence registration and text searches within automatically archived documents in the Infopoint software. The objective is to achieve greater efficiency in these processes and, through learning, establish intelligent document management.

2. Background

IMG Procesos y Tecnología SAS (IMG) is a Colombian company specializing in software development and the creator of Infopoint, an electronic document management system (EDMS) that oversees the entire lifecycle of a company's documents. It includes creation or receipt, workflow management, and organized storage in a centralized repository for preservation or eventual destruction. Infopoint is software for the management, organization, traceability, and security of both physical and electronic documents.

Infopoint includes various modules: e-correspondence (incoming correspondence registration through internal and external dynamic forms for physical and electronic correspondence), e-workflow (workflow management), integrations (services for system integration to ensure interoperability with Infopoint), e-postal office (sending responses and documents via email), e-archive (repository for physical documents with precise location information in physical warehouses), document center (centralized repository for electronic and digitized physical documents, with multiple classification criteria and search functions), e-reports and indicators (generation of document management reports), and e-admin (system administrator).

Some features of the system include:

- Registration of incoming correspondence or document production from various sources such as physical correspondence, emails, internal and external dynamic forms for physical and electronic correspondence, etc.
- Design and management of processes and workflows based on documents to organize and control work methods and improve response times.
- Systematization of various document management processes such as complaints, legal actions, procedures, administrative and financial processes, etc., both internal and external.

¹LSTM: Abbreviation for "long-short term memory," which is an extension of neural networks that enhances their memory to learn from experiences, remembering inputs both during the training process and in their usage over an extended period, allowing this to be used as input to refine results during execution.

²Neural network: type of machine learning process. An artificial intelligence method that teaches a system to process data in a way inspired by how the human brain operates. It uses objects known as neurons, which are interconnected in a layered structure, creating an adaptable system that can learn from its mistakes and develop continuous improvement.

³Convolutional: type of error-detection code where each m-bit information symbol is transformed, when encoded, into an n-bit symbol, where m/n is the code rate ($n \geq m$).

- Review and approval of outgoing correspondence with the option of electronic or digital signatures.
- Management and control of loans for physical documents with notifications of due dates.
- Standardized structuring for archiving each document with multiple query criteria.
- Parameterization of access permissions and actions for users.
- Traceability of each action performed.

The organization of electronic documents in Infopoint, and in general, in any electronic document management system (EDMS), requires a specific working methodology, explained in the following steps:

1. Create a logical folder structure: A folder structure reflecting the organizational structure must be designed by creating records retention tables (RRT), which are designed to assist companies in efficiently managing their documents by determining retention and disposal periods for the documents.
2. Establish metadata: Metadata should be employed to classify and label documents to be easily identified and retrieved.
3. Apply a naming policy: Clear rules must be defined for naming documents to maintain order and facilitate their retrieval.
4. Preserve the original order: When importing documents into the EDMS, companies must ensure the preservation of the original order as much as possible. It entails maintaining the folder structure and file names as they are in their original location unless it is necessary to reorganize them.
5. Record changes and movements: The EDMS should audit any changes or movements made to the documents, including information about who, when, and why the change was made. Maintaining an audit trail ensures traceability and transparency in document management.
6. Implement access controls: The EDMS should allow for the establishment of appropriate access permissions to protect the confidentiality and integrity of documents; only authorized personnel should be able to view, modify, or delete documents as necessary.
7. Train staff: Proper training should be provided to staff on the use of the EDMS following established policies and procedures, which should include how to search, retrieve, modify, and archive documents correctly.

By following the above steps, companies can effectively manage electronic documents without compromising the principles of provenance and original order.

3. Literature review

Automation of document management has been the subject of numerous studies and research. Similar to the automation in the Infopoint document management software, automation is applicable in library processes due to the need to manage a large volume of documentary material.

Library automation is a process involving the application of technologies to enhance services and information management in libraries. The benefits of library automation include increased efficiency, improved service quality, greater access to information, greater data security, and increased flexibility (Arriola; Montes de Oca, 2014).

Specifically, automation is being utilized in libraries to provide the following services: online catalogs (enabling users to quickly and easily search for materials in libraries), bibliographic databases (access to journal articles, books, and other materials), interlibrary loan services (borrowing materials from other libraries), online reference services (remote access to librarian assistance), and technology training services (helping users learn to use new technologies) (Bibliopos, 2023).

The Sistema Integral de Automatización de Bibliotecas (Comprehensive Library Automation System, SIAB) is a system that aids libraries in automating their processes and services. A survey on the automation

of services was conducted in university libraries in the metropolitan area of Mexico City, obtaining interesting results. Most libraries surveyed have a SIAB in place, with the most commonly used systems being Siabuc, Aleph, and Logicat. However, the implementation level of these systems is low, primarily due to a lack of training for the personnel in charge. Automation has often been limited to managing the online catalog and book lending (Arriola; Montes de Oca, 2014). Another significant finding is that the cost of the system and high maintenance expenses are the main obstacles to increased library automation. Furthermore, only one library had installed open-source software, indicating a lack of awareness about such alternatives.

In general, the study results indicate that university libraries in the metropolitan area of Mexico City still have a relatively low level of automation. To improve the situation, libraries must invest in staff training, reduce maintenance costs, and explore open-source software alternatives.

Advances in automation in document management systems are a growing trend in businesses, with various identified benefits such as precision in operations and workflow, organized recordkeeping, and improvement in delivery times (Smart Government Solutions, 2021). Furthermore, document management automation enhances efficiency and ensures companies' regulatory compliance, provides greater security by protecting documents, and improves coordination among departments and employees in a company (Prevecionar, 2023).

Digital repositories need effective and efficient retrieval of stored material. The application of machine learning techniques and first-order logic in the processing of electronic documents is crucial for design identification, content classification, and text extraction. These technologies play an essential role in document management in digital environments and have the potential to significantly improve efficiency and accuracy in this constantly evolving field (Esposito *et al.*, 2005).

Jayoma *et al.* (2020) conducted a study on the daily records produced by the Department of Social Work and Development in Caraga, Philippines, emphasizing that their conventional record management system makes it difficult to track and maintain records. The authors demonstrated the automation of record classification using the Tesseract⁴ library to recognize and extract text, simplifying the tasks of classification, indexing, and archiving of records. With the help of this system, the custody and retrieval of records are simplified, facilitating the work of record officers.

Efficient knowledge management depends on the smooth integration of information from both digital and paper documents. A practical approach involves digitizing documents from natural images, requiring precise location within the image. Traditional methods fall short in the face of extreme variations in perspective and background. Leveraging the robustness of deep convolutional neural networks (CNN), Javed and Shafait (2017) propose an innovative method based on CNN for real-time, precise document location, addressing location as a key point detection challenge, predicting the four corners of the document jointly and continually refining the results.

Automatic document categorization is crucial to processing a large amount of explicit knowledge documents in an organized manner. A document classification methodology based on neural network technology is proposed by Trappey *et al.* (2005). Key phrases were extracted from the document set through text processing, and the meaning of these key phrases was determined based on their frequency of appearance. A keyword correlation analysis model was then applied to calculate the similarity between key phrases, and synonyms of highly similar terms were extracted. The backpropagation neural network model was adopted as the classifier. The result was the identification of the appropriate category of a document based on the hierarchical classification scheme of documents, in this case, the international patent classification standard. In the prototype system, the automatic classification module helps users classify patent documents, and the search

⁴Tesseract: OCR engine based on a neural network (LSTM) focusing online and character recognition.

module helps users quickly find the correct patent documents. The result shows a significant improvement in the classification and identification of documents in explicit knowledge management.

In a study, Mukherjee and Roy (2021) present a methodology that facilitates the electronic editing and search of degraded documents by digitizing degraded Bengali documents using optical character recognition (OCR)⁵ technology and then achieving classification of a wide variety of characters using a convolutional neural network model. Based on the literature, the methodology to be discussed is proposed below.

4. Methodology

The proposed methodology for the development of the project aimed at automation functionalities in the Infopoint document management system to generate “intelligent document management” is based on four development phases, as illustrated in Figure 1:

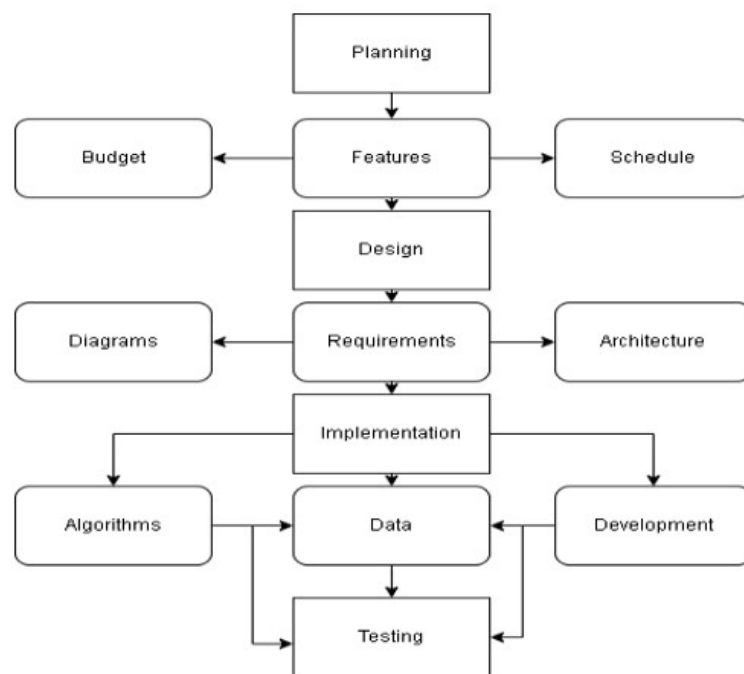


Figure 1. Proposed methodology

Source: own elaboration.

The project scope was defined in the planning phase, objectives were outlined, and a schedule with products, timelines, and activities was established. Execution times and resources (human, infrastructure, and technical) were allocated. Deliverables for each activity and responsible parties were assigned.

The design phase focused on identifying Infopoint functionalities that could be addressed to reduce manual levels in operations and document searches. Additionally, the development of the requirements document, technology selection, and architecture creation were undertaken.

The implementation phase involves the development of algorithms and data management based on functionalities identified in the design phase and outlined in the requirements document. For this case, Infopoint’s programming language is .NET - C#, and technologies such as Aspose.PDF, OCR, neural network, long-short term memory (LSTM), convolutional, and Tesseract are necessary to extract data from digitized

⁵OCR: Abbreviation for optical character recognition, a process by which an image of text becomes into a text format that can be read and interpreted by systems to perform tasks such as text editing, searching, word counting, etc.

documents, process the data, transform it into information, and use it as input for a neural network that identifies the form that corresponds to the document. It allows for the automatic completion of fields configured for automation. Database architecture was also updated.

The final phase encompasses the functional evaluation of the development, algorithms, and data management against defined requirements, obtained results, implementation conclusions, and testing.

4.1. Planning

For the development of automation functionalities in Infopoint, it was initially identified that the correspondence registration process involves a high level of manual intervention. Therefore, the automation in Infopoint was specifically focused on this incoming correspondence functionality to achieve significant time savings and reduce manual and repetitive tasks for system users. The process functioned before automation, as depicted in Figure 2.

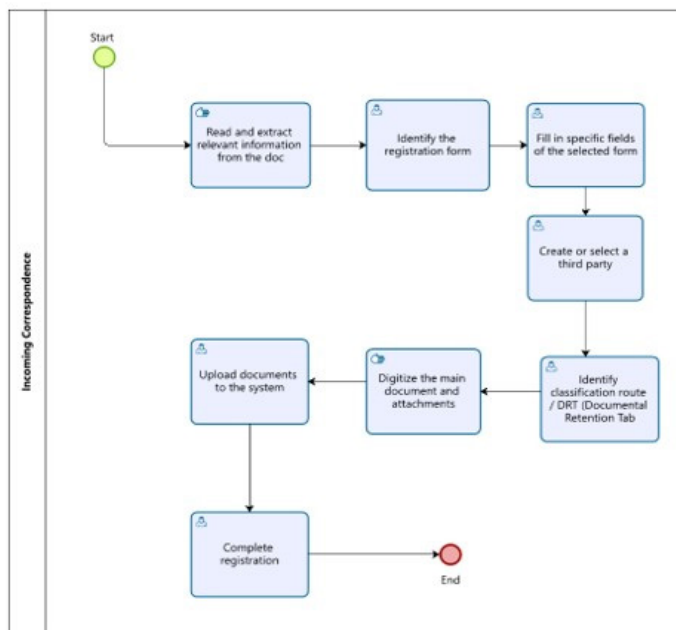


Figure 2. Incoming correspondence functionality in Infopoint before automation
Source: own elaboration.

Each step in the process is marked with an icon indicating whether it is a manual process, external to the Infopoint system, or an internal system process, as follows:

- Manual (🖱️): Manual tasks performed by a person outside the Infopoint system.
- User (👤): Manual tasks performed by a person within the system.

In total, there were eight manual tasks in this process.

4.2. Design

To initiate the design process, research was conducted to validate how artificial intelligence could serve to replace two steps of the inbound correspondence filing process functionality in Infopoint, as shown in Figure 2:

- Reading and extracting relevant information from the document
- Identifying the registration form.

For the step of reading and extracting relevant information from the document, it was identified that to apply artificial intelligence, it is necessary to employ image processing techniques by performing a series of sequential actions such as image preprocessing, text region detection, and feature extraction for information classification and text recognition. These actions are described in Figure 3.



Figure 3. Steps to be carried out for reading and extracting text from images.

Source: own elaboration.

The first step in the image preprocessing process aims to enhance the image quality for subsequent information extraction. To achieve this, various filters are applied based on the specific need to be addressed. These filters may include noise reduction, contrast adjustments, perspective correction, and size normalization, among others. Once the image has been filtered, region detection is performed. This process seeks to identify and segment areas of the image containing relevant information using contour detection algorithms. With the regions identified, feature extraction is carried out. In the specific case of text character detection, convolutional neural networks (CNN) are utilized. These networks are designed to identify text and are trained on pre-existing data containing examples of different characters in each language, represented in pixel⁶ intensities. Each character identification processed by the neural network provides the representation of the detected character and the associated certainty percentage with the identification process. The use of different types of neural networks depends on the particular problem being addressed. For the specific case of Infopoint, where the goal is to read and extract information from a document, the most viable approach was utilizing convolutional neural networks, as they are widely employed for image processing due to their learning and feature extraction capabilities.

In Infopoint, where character identification is performed on document images, filters were applied to preprocess the image before sending it to the neural network, employing a combination of techniques including grayscale image transformation and Gaussian⁷ smoothing filter.

⁶Pixel: The smallest homogeneous unit in color that is part of a digital image.

⁷Gaussian smoothing (Suavizado gaussiano): Filter used on data to eliminate noise and other unwanted artifacts to enhance its quality and facilitate analysis. Widely used technique in image processing.



Figure 4. The result of applying a combination of preprocessing filters to the original image
Source: own elaboration.

In Figure 4, a change to grayscale is applied to the original image to standardize and simplify the information to be sent to the neural network by reducing it to a single channel (grayscale) to facilitate its analysis. Similarly, noise reduction is employed to eliminate possible imperfections that may occur in the digitized document and to remove pixel intensity variations by applying the Gaussian smoothing filter.

As an input parameter, the preprocessed image was sent, and upon processing through the neural network, three components were obtained in the output:

- *Confidence percentage (Conf)*: Percentage of certainty generated by the neural network that the output text is correct.
- *Output text (Text)*: Text output according to the neural network process.
- *Final confidence percentage*: Total percentage of certainty in text extraction from the image.

RESPONSE OF THE NEURAL NETWORK AFTER IDENTIFICATION OF TEXT IN THE IMAGE

INPUT										OUTPUT	
Level	Page_num	Block_num	Par_num	Line_num	Word_num	Left	Top	Width	Height	Conf	Text
1	2	0	0	0	0	0	0	139	194	-1	
2	2	1	0	0	0	33	0	75	71	-1	
3	2	1	1	0	0	33	0	75	71	-1	
4	2	1	1	1	0	33	0	75	71	-1	
5	2	1	1	1	1	33	0	75	71	95.000000	
2	2	2	0	0	0	13	93	121	47	-1	
3	2	2	1	0	0	13	93	121	47	-1	
4	2	2	1	1	0	13	93	121	15	-1	
5	2	2	1	1	1	13	93	121	15	92.598160	detecciónde
4	2	2	1	2	0	19	112	109	12	-1	
5	2	2	1	2	1	19	112	109	12	92.521561	regionesde
4	2	2	1	3	0	45	129	56	11	-1	
5	2	2	1	3	1	45	129	56	11	96.766953	texto
2	2	3	0	0	0	0	168	138	12	-1	
3	2	3	1	0	0	0	168	138	12	-1	
4	2	3	1	1	0	0	168	138	12	-1	
5	2	3	1	1	1	0	168	63	12	96.661247	técnicas
5	2	3	1	1	2	69	168	16	12	96.450630	de
5	2	3	1	1	3	91	168	47	12	96.782455	umbral

94.00%

Figure 5. The response generated by the neural network when the preprocessed image is sent as an input parameter
Source: own elaboration.

Once the neural network execution is performed, the results observed in Figure 5 are obtained. The experimentation conducted with the convolutional neural network (CNN) used by the Tesseract library, the optical character recognition (OCR) engine employed for extracting text from digitized documents, yielded significant results highlighting the importance of image preprocessing in the performance of these models. Initially, an experiment was conducted by sending an image without any preprocessing (without any applied filters) directly as input to the neural network. It was observed that the network extracted the word “técnicas” with a certainty percentage of 44 %. It indicates a lack of information for the neural network or the necessity of adjusting input parameters to enhance its response. Upon applying preprocessing to the image, the neural

network achieved a 94 % certainty in text extraction (Figure 5). This performance improvement of the CNN underscores the importance of considering data preprocessing in Infopoint during the development phases of the component to optimize the response from the machine learning model, enhancing predictive capability and model accuracy.

Among the natural language algorithms employed for understanding and analyzing human language, the Tesseract library was utilized, a library with a database of training files in multiple languages, which offers precision and versatility when applied to the information extraction task. When Tesseract is applied for character identification in images of digitized documents, it is implemented in conjunction with the natural language processing algorithm of tokenization to store this information in a structured manner in a database. Tokenization involves dividing the text into smaller units, such as words or phrases, to facilitate storage and subsequent analysis.

This information structuring from the extraction process enables the implementation of optimized search functionalities on Infopoint, facilitating access and retrieval of documents based on their content. Additionally, it allows the utilization of classification algorithms to automatically assign the corresponding completion form to the digitized document based on its tokenized content. One of these algorithms is XGBoost, a supervised machine learning algorithm based on decision trees, known for its effectiveness and accuracy in classification tasks. Upon analyzing the feasibility of applying such algorithms, it is determined, based on the problem characteristics and advantages of each of these algorithms, to employ XGBoost, which carries out a series of activities (such as variable definition, data collection, decision trees as base models, and regularization and feature handling) to obtain the classification result.

The selection of tasks to automate was based on the capability of process automation technology or RPA⁸ (robotic process automation). Figure 6 illustrates the workflow of incoming correspondence after automation.

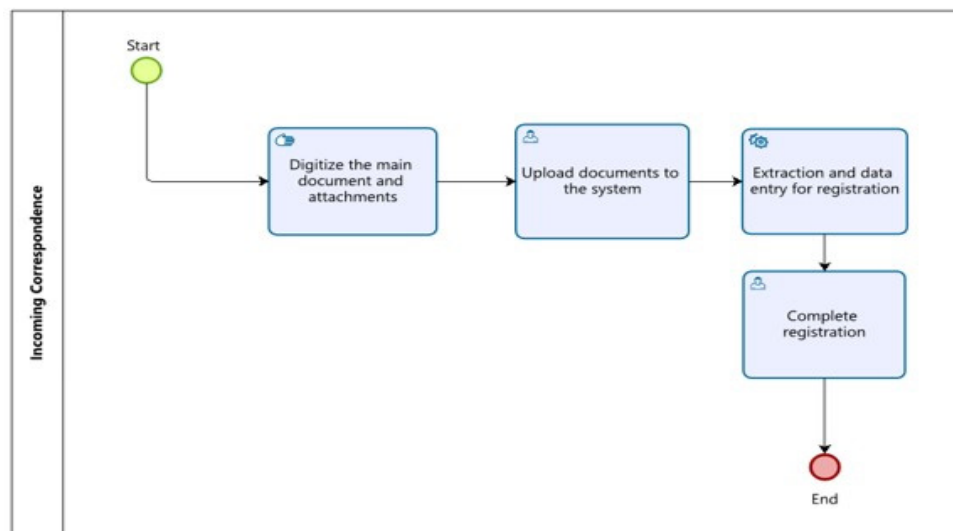



Figure 6. Incoming correspondence functionality in Infopoint after automation

Source: own elaboration.

⁸Process automation (or robotic process automation, RPA) involves optimizing tasks or activities through software. The tasks targeted for automation are typically manual, repetitive, and easy-to-perform activities.

The icon  represents a task automatically performed by the system.

Following automation, only three tasks (external or internal to the system) need to be carried out manually, while the remaining five of the initial flow are automatically performed by the system and synthesized in the extraction and completion of information for registration.

4.3. Implementation

Multiple data extraction technologies were utilized for digitized documents to achieve automation of the incoming correspondence functionality in Infopoint,x. Subsequently, data was processed, transformed into information, and used as input for a neural network that identifies the form corresponding to the document. It allows for the automatic completion of configurable fields, aiming to make the correspondence entry process more efficient.

Figure 7 presents the process used for document classification, observing, and explaining the steps of the preparation, training, and production phases. The process consists of five steps: form design, model definition, model training, automated filing process, and auto-filling of form fields.

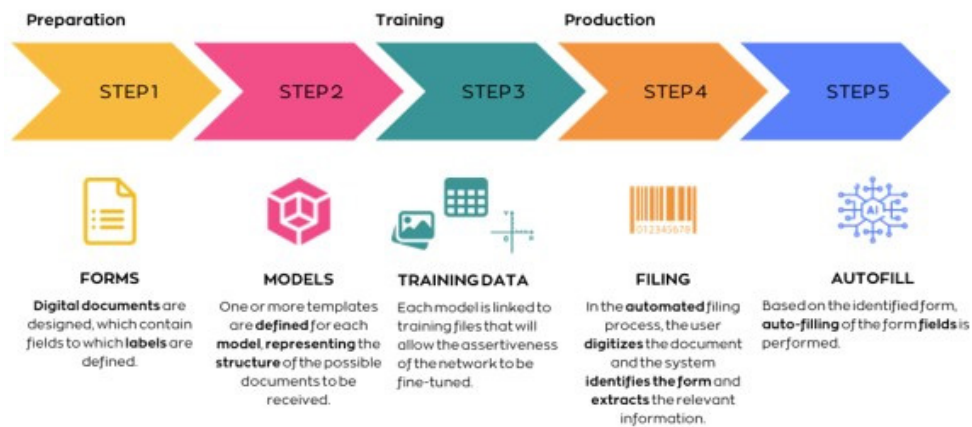


Figure 7. Process for document classification
Source: own elaboration.

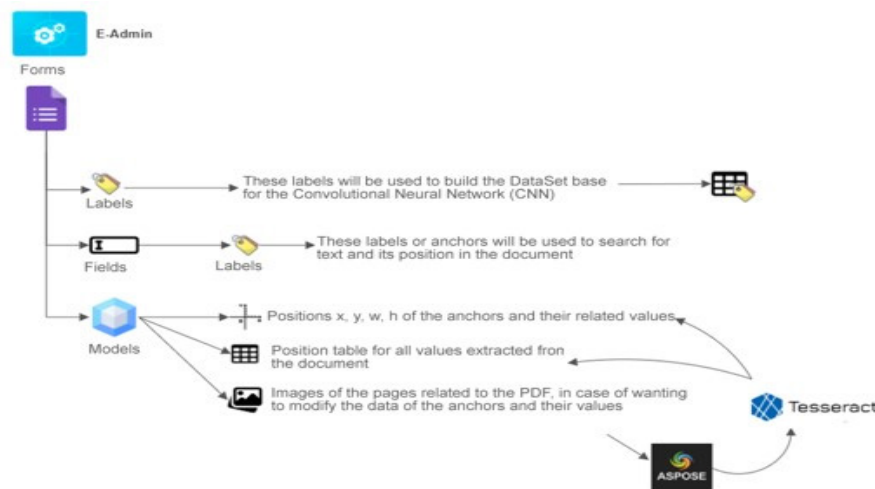









Figure 8. Technical model. Configuration of forms in the e-correspondence module
Source: own elaboration.

In Figure 8, the Forms () represent digital documents where users input data in a structured manner into configured fields () based on chosen data types and defined requirements. The system can then store and process the data, converting it into information. Two new components were included in this module:

- **Labels** (): Words or composite texts are used for searching within the document, acting as keywords and identification mechanisms for form validation logic. These were added for both the form and the configured fields within it.
- **Models** (): Objects serve as guidelines for the system to be reproduced or copied once the form is “published” for use in e-correspondence. Models consist of the following information:
 - o (): Pixel coordinates⁹ of field labels and their corresponding values within the model document.
 - o (): Table of values for all information extracted from the document.
 - o (): Set of images representing the .JPG version of each page composing the model document.

A document imported through the system by an administrative user served as the basis to build a model. This document underwent processing through the Aspose.PDF¹⁰ library to obtain images of its pages. Subsequently, these images were transformed in Infopoint using the Tesseract library.

Once the models were created and configured, the system was allowed to use the information obtained to initiate training of the neural network responsible for identifying the form needed for automatic completion. It was based on the document digitized in the e-correspondence module.

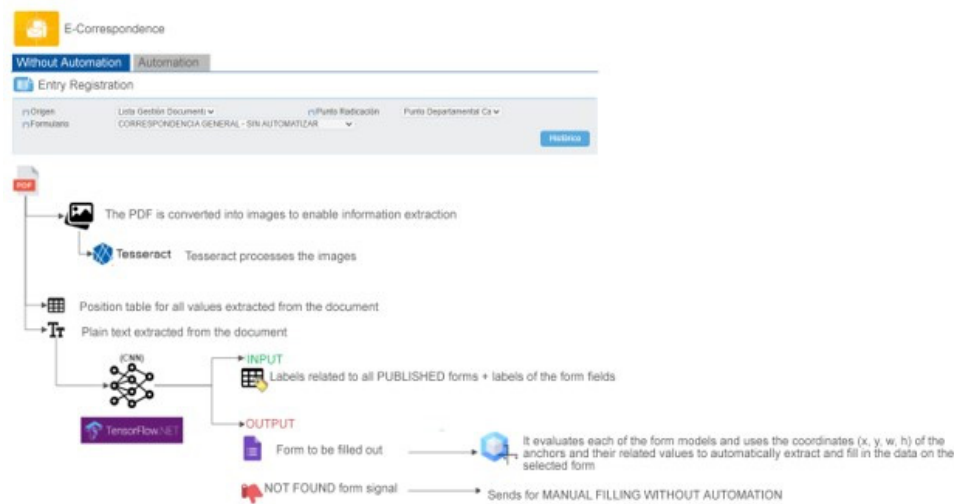


Figure 9. Technical model. Functionality in the e-correspondence module

Source: own elaboration.

Figure 9 shows that the “automation” option must be selected during correspondence registration within e-correspondence. Upon importing a digitized PDF file, images representing each page of the document are extracted. Subsequently, the images are processed through the Tesseract library to obtain two elements:

- Table of positions of values for each page of the document
- Plain text extracted from all pages of the document.

⁹Coordinate: A value system used to determine the position of a point on a plane.

¹⁰Aspose.PDF: High-level API for document manipulation with the .PDF extension. API: Abbreviation for “application programming interfaces.” It is a set of definitions and protocols used to develop and integrate applications, enabling communication through a set of rules.

The plain text element is analyzed by the convolutional neural network (CNN) developed using TensorFlow technology with a specialized .NET library called TensorFlow.NET. This library provides access to a set of machine-learning tools using the existing TensorFlow library as the underlying engine.

The convolutional neural network (CNN) was chosen because it allows it to base itself on training data, marked in Figure 9 as input, represented by labels related to the form and those related to its fields. It enables the neural network to extract certain features for document classification and form identification automatically.

The output of the neural network consists of two pieces of data:

- Form name that the neural network identifies to be filled out, based on its machine learning.
- The numeric value, known as the confidence level, indicates, on a scale of 0 to 100, the degree of certainty or confidence with which the neural network, based on its training, produces the result.

The system determines if the result is reliable based on the confidence level obtained. If so, it loads the resulting form in the browser and, using the models automatically extracts the related values for each field of the document (identified with labels related to the fields of the form in the previous parameterization). It then automatically fills in the corresponding fields in the digital form, partially automating the correspondence registration process.

In cases where the confidence level is unreliable, the system generates an alert informing the user that it was unable to find the form and automatically redirects to the “without automation” functionality.

In summary, Figure 10 illustrates and explains the steps from filing to auto-filling in the classification and automation process.

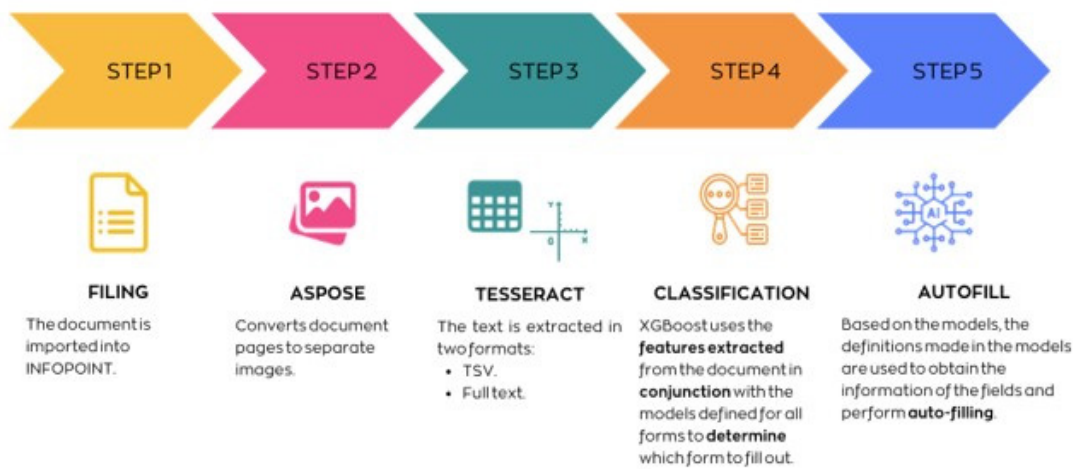


Figure 10. Summary of the classification and automation process
Source: own elaboration.

Following automation, Infopoint can perform text searches within PDF documents that were automatically registered through the e-correspondence module. To achieve this, intelligent document processing (IDP) was employed. IDP converts unstructured and semi-structured data in a document into usable structured information. Using machine learning and OCR, the required data is extracted from the documents, and IDP classifies these documents based on the extracted data, utilizing automation services to store the documents in the Infopoint document center.

Through a unique search field implemented in the document center, it is possible to search for documents processed with IDP technology. The search is conducted across all extracted data from different

documents, displaying documents, registrations, folders, and records related to any extracted data that was the subject of the search. It optimizes searches within the documents archived in the system, enabling quick data queries within large volumes of documents.

The text search within documents is a process that comprises five general steps: query input, document search, coincidence, classification, and presentation of results. The details of these steps are in Figure 11 below:



Figure 11. Document search process

Source: own elaboration.

The process of extracting information contained in files that have been registered through automation is illustrated in Figure 12:

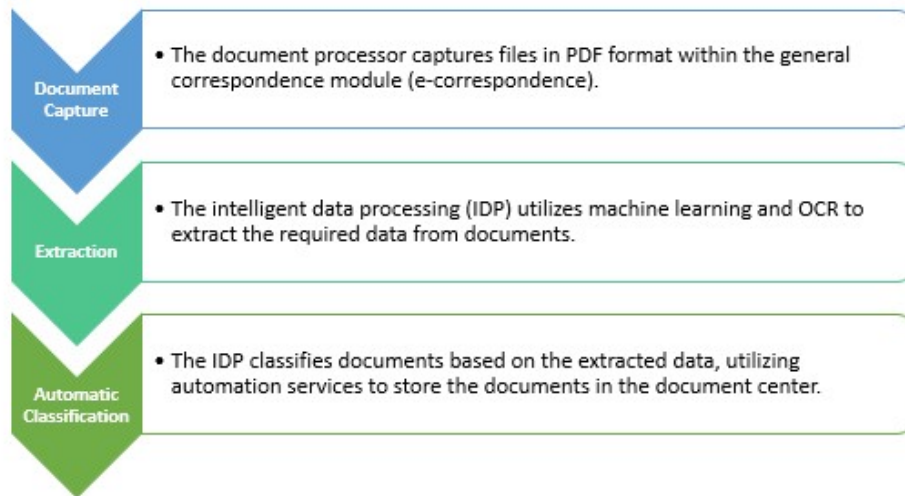


Figure 12. Information extraction process from files through automation in their registration

Source: own elaboration.

Once the described process is completed for a document, it is saved in the Infopoint document center module and becomes available for retrieval whenever the user needs it.

To perform the document retrieval process in the document center, the search is conducted only on files that have undergone the automation process in the correspondence module, selecting the “automation” option.

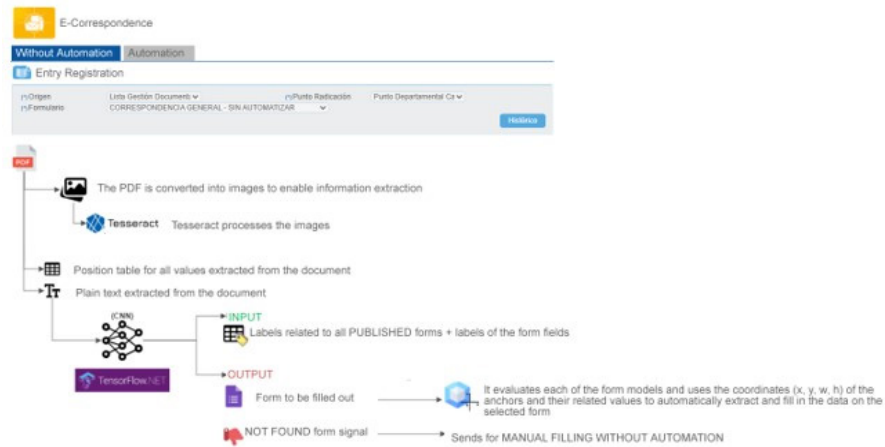


Figure 13. Technical model of the functionality to be performed in the e-correspondence module
Source: own elaboration.

As detailed in Figure 13, the process begins with the import of digitized PDF files, from which an image representation of each page is extracted to be processed with the Tesseract library.

As mentioned before, by using Tesseract in Infopoint as OCR technology, documents containing pre-trained data (language model and grammar rules) are used to identify patterns and characteristics of characters through shape recognition algorithms and neural network use for machine learning, resulting in two key elements:

- (📄) Position table of values for each page (see Figure 14): This table contains the following information, one for each element found on the pages:
 - o Hierarchical level of abstraction of the object, with values for:
 - 1: Page
 - 2: Block
 - 3: Paragraph
 - 4: Line
 - 5: Word
 - o Page number of the object
 - o Block number
 - o Paragraph number
 - o Line number
 - o Word number
 - o “x” coordinate in pixels of the top-left corner of the text bounding box, starting from the left of the image
 - o “y” coordinate in pixels of the top-left corner of the text bounding box, starting from the top of the image
 - o Width of the text bounding box in pixels
 - o Height of the text bounding box in pixels
 - o Confidence value: -1 for levels 1 to 4. For level 5: values between 0 (no confidence) and 100 (maximum confidence)
 - o Detected text: empty for all levels except 5
- Plain text extracted from all pages: Translation into digital text contained in the document.

to find relevant results. By entering a simple or compound word in the search field of the document center, the system employs this text index to retrieve documents containing the requested information and returns them to the user as the result of the query.

Searches are an essential part of document management since, after saving a document, the expectation is to be able to quickly search and find it. A single search field was implemented in the document center to facilitate this, allowing searches for documents processed with IDP technology. The search is conducted across all data extracted from different documents, displaying documents, registrations, folders, and records related to any extracted data in the search result.

5.1. Evaluation

Ten data entry operators or inputters were involved in testing the development of the incoming correspondence automation. They processed ten registrations for three types of forms, first manually and then using the automation functionality. The registration forms were for:

- General correspondence
- Invoices
- Constitutional actions

The time duration for each manual registration was measured, and the same procedure was repeated for automated registrations.

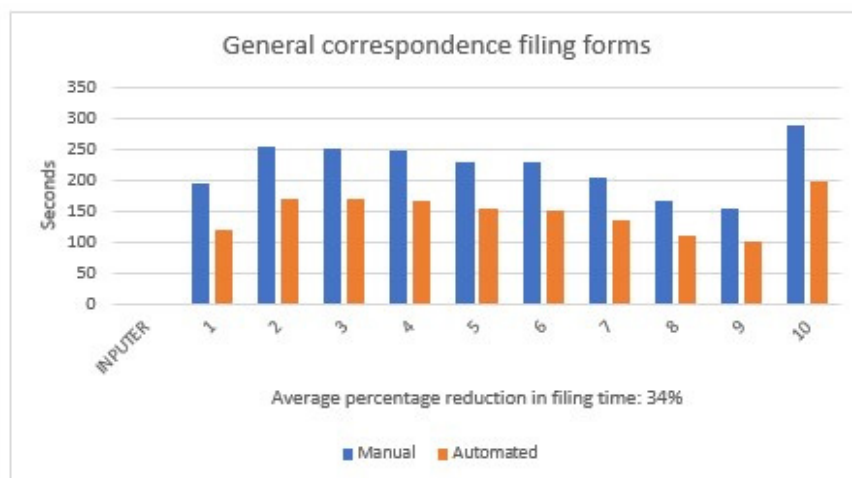


Figure 15. Duration of general correspondence registration. Manual vs. automated
Source: own elaboration.

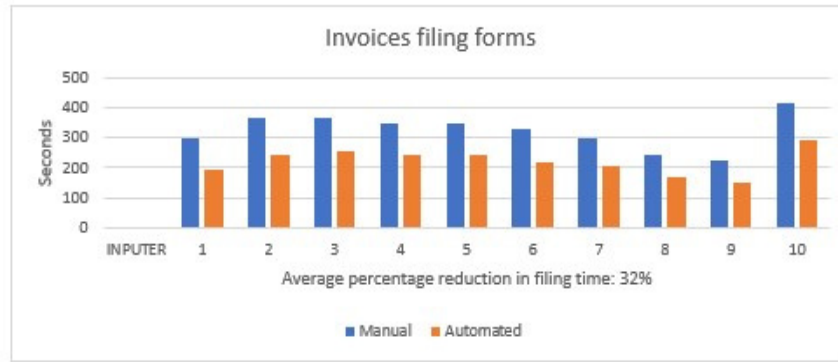


Figure 16. Duration of invoice registration. Manual vs. automated
Source: own elaboration.

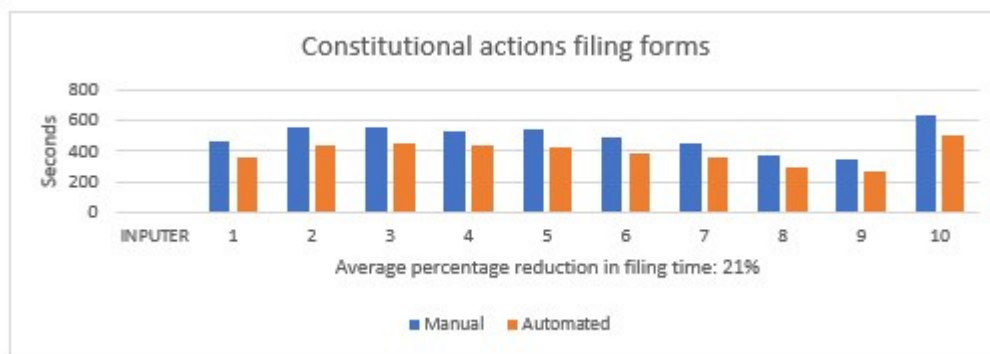


Figure 17. Duration of constitutional actions registration. Manual vs. automated
Source: own elaboration.

As seen in Figures 15, 16, and 17, a significant reduction in the registration time was observed for each type of registration when comparing manual versus automated registrations, resulting in the following results:

- General correspondence: 34 %
- Invoices: 33 %
- Constitutional actions: 21 %

This reduction is considered highly significant for an entity as it represents more than a substantial percentage of an employee's time. This time can be utilized for other activities that can contribute greater value to the organization.

Implementing automation technologies has streamlined and optimized the incoming correspondence process in Infopoint. An intelligent system was developed, capable of automatically classifying incoming documents, assigning relevant tags, and directing them to the corresponding departments within the organization. It has drastically reduced the time spent on manual document classification and improved operational efficiency. Additionally, automated workflow processes were implemented to ensure smooth correspondence management, allowing a quicker and more accurate response to client requests.

Regarding the search and text extraction from files, a search was conducted on ten documents using both the keywords criterion (metadata related by the user to the document) and the intelligent processing panel (automated registration). The average time for each type of search is shown in Table 1 below:

Table 1. Search by metadata vs. search by intelligent processing

General correspondence Filing Form				
Documento	Keywords	Intelligent processing		
	Averag Time (Seconds)	Averag Time (Seconds)	Averag Time Reduction	Reduction percentage
1	10.00	8.00	2.00	20%
2	6.00	3.00	3.00	50%
3	4.00	2.00	2.00	50%
4	3.00	1.50	1.50	50%
5	9.00	3.40	5.60	62%
6	30.00	12.50	17.50	58%
7	20.00	13.50	6.50	33%
8	5.00	4.20	0.80	16%
9	8.00	6.00	2.00	25%
10	4.00	2.00	2.00	50%
Average reduction percentage				41%

Source: own elaboration.

During the measurement of response times for a total of ten documents, an average reduction in search time of 41 % was observed.

6. Conclusions

Implementing automation technologies has enabled intelligent and efficient document management, improving the development of users’ business needs.

The results demonstrate the tangible benefits of implementing advanced technologies in document management. The automation of incoming correspondence functionalities has enhanced operational efficiency and accelerated client response times. Furthermore, intelligent text and content searches have allowed faster and more precise retrieval of relevant information.

Implementing intelligent and efficient document management impacts productivity, decision-making, and organizational responsiveness. In addition, reducing paper usage and promoting digitization encourages sustainability, contributing to environmental protection.

The project has laid the groundwork for future improvements and developments in electronic document management. Implementing cutting-edge technologies has allowed users to enjoy more efficient, accurate, and agile document management.

References

- Archivo General de la Nación (2013). *Gestión Documental*.
<https://glosario.archivogeneral.gov.co/vocab/index.php?tema=142#:~:text=Conjunto%20de%20actividades%20administrativas%20y,facilitar%20su%20utilizaci%C3%B3n%20y%20conservaci%C3%B3n>
- Arriola, Óscar; Montes de Oca, Evangelina (2014). Sistemas Integrales de Automatización de Bibliotecas: una descripción sucinta. *Bibliotecas y Archivos*, 1(4), 47-70.
- Bibliopos (2023). *Sistemas integrados de automatización de bibliotecas. Situación actual y tendencias de futuro*.
<https://www.bibliopos.es/Biblion-A2-Biblioteconomia/23Sistemas-Integrados-Automatizacion-Bibliotecas.pdf>
- Esposito, Floriana; Ferilli, Stefano; Basile, Teresa; di Mauro, Nicola (2005). Intelligent document processing. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (Vol. 2, pp. 1100-1104). Institute of Electrical and Electronics Engineers.
<https://www.doi.org/10.1109/ICDAR.2005.144>
- Javed, Khurram; Shafait, Faisal (2017). Real-Time Document Localization in Natural Images by Recursive Application of a CNN. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 105-110). Institute of Electrical and Electronics Engineers.
<https://www.doi.org/10.1109/ICDAR.2017.26>
- Jayoma, Jaymer; Moyon, Elbert; Morales, Edsel (2020). OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines. In *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)* (pp. 1-6). Institute of Electrical and Electronics Engineers.
<https://www.doi.org/10.1109/HNICEM51456.2020.9400000>
- Kantan Software (2022, May 15). *4 Costes ocultos de un control de documentos deficiente*.
<https://www.kantansoftware.com/blog/4-costes-ocultos-de-un-control-de-documentos-deficiente/>
- Llamas, Jonathan; López, José (2020, October 2). Automatización de procesos. *Economipedia*.
<https://economipedia.com/definiciones/automatizacion-de-procesos.html>
- Mukherjee, Jayati; Roy, Utpal (2021) Recognition of Degraded Bangla Documents Using Hybrid Deep Neural Network Model. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 254-259). Institute of Electrical and Electronics Engineers.
<https://www.doi.org/10.1109/ICACITE51222.2021.9404691>
- Prevencionar (2023, May 30). *Beneficios de la automatización en la gestión documental CAE: Eficiencia, cumplimiento y coordinación efectiva*.
<https://prevencionar.com/2023/05/30/beneficios-de-la-automatizacion-en-la-gestion-documental-cae-eficiencia-cumplimiento-y-coordinacion-efectiva/>

- Rangel, Erika (2017). *Guía de implementación de un sistema de gestión de documentos electrónicos de archivo-SGDEA. Archivo General de la Nación.*
https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicacionees/ImplementacionSGDEA.pdf
- Ricketts, Whitney (2014, March 27). *Inaugural Creative Jobs Report Reveals New American Dream. Creative Live.*
<https://www.creativelive.com/blog/creative-jobs-report/>
- Smart Government Solutions (2021, December 20). *Automatización de procesos en los sistemas de gestión documental. Smart Government Solutions.*
<https://smartgs.com.mx/automatizacion-de-procesos-en-los-sistemas-de-gestion-documental/>
- Storecheck (2022, May 6). *10 beneficios de automatizar tu operación.*
<https://blog.storecheck.com.mx/10-beneficios-de-automatizar-tu-operacion/>
- Trappey, Amy; Lin, Simon; Wang, Albert (2005). Using neural network categorization method to develop an innovative knowledge management technology for patent document classification. In *Proceedings of the Ninth International Conference on Computer Supported Cooperative Work in Design* (Vol. 2, pp. 830-835). Institute of Electrical and Electronics Engineers.
<https://www.doi.org/10.1109/CSCWD.2005.194293>