information

MDPI

Article

An Automated Repository for the Efficient Management of Complex Documentation

José Frade [1,*] and Mário Antunes [1,2,*]

[1] School of Technology and Management, Polytechnic University of Leiria, 2411-901 Leiria, Portugal
[2] INESC TEC, CRACS, 4200-465 Porto, Portugal
* Correspondence: 2220723@my.ipleiria.pt (J.F.); mario.antunes@ipleiria.pt (M.A.)

Abstract: The accelerating digitalization of the public and private sectors has made information technologies (IT) indispensable in modern life. As services shift to digital platforms and technologies expand across industries, the complexity of legal, regulatory, and technical requirement documentation is growing rapidly. This increase presents significant challenges in managing, gathering, and analyzing documents, as their dispersion across various repositories and formats hinders accessibility and efficient processing. This paper presents the development of an automated repository designed to streamline the collection, classification, and analysis of cybersecurity-related documents. By harnessing the capabilities of natural language processing (NLP) models—specifically Generative Pre-Trained Transformer (GPT) technologies—the system automates text ingestion, extraction, and summarization, providing users with visual tools and organized insights into large volumes of data. The repository facilitates the efficient management of evolving cybersecurity documentation, addressing issues of accessibility, complexity, and time constraints. This paper explores the potential applications of NLP in cybersecurity documentation management and highlights the advantages of integrating automated repositories equipped with visualization and search tools. By focusing on legal documents and technical guidelines from Portugal and the European Union (EU), this applied research seeks to enhance cybersecurity governance, streamline document retrieval, and deliver actionable insights to professionals. Ultimately, the goal is to develop a scalable, adaptable platform capable of extending beyond cybersecurity to serve other industries that rely on the effective management of complex documentation.

Keywords: cybersecurity; automated repositories; document management; compliance and legal documentation; document categorization; large language models

Academic Editors: Tudor Groza, Junwen Duan and Fangfang Li

Received: 7 February 2025
Revised: 27 February 2025
Accepted: 2 March 2025
Published: 5 March 2025

Citation: Frade, J.; Antunes, M. An Automated Repository for the Efficient Management of Complex Documentation. *Information* **2025**, *16*, 205. https://doi.org/10.3390/info16030205

1. Introduction

The digitalization of the public and private sectors is accelerating, with information technologies (IT) becoming an essential part of daily life. Services are rapidly shifting to digital platforms, and emerging technologies such as artificial intelligence (AI) are finding applications in various fields [1]. Although digital transformation offers enhanced accessibility and efficiency, it also introduces considerable risks. The increasing value of IT assets and personal data attracts malicious actors, raising concerns about privacy abuses and data misuse by companies and governments [2–4]. As society becomes more dependent on digital systems, ensuring the quality, reliability, and security of critical infrastructure, particularly in sectors such as healthcare and energy, remains a pressing challenge [5].

To address these challenges, governments and international institutions have issued a growing number of regulations, legal documents, and technical guidelines aimed at

guiding organizations through their digital transformation. In addition, internationally recognized institutions provide technical standards and guidelines that, while not legally binding, offer crucial measures and policies. Implementing these recommendations can improve resilience to cyber threats, improve cyber policies, and strengthen management practices. In many cases, adherence to such standards means a competitive advantage and is required to obtain certification or formal recognition.

Despite their availability on various online platforms, the management of these documents is becoming increasingly complex due to the high volume of publications delivered on a daily basis, the diversity of repositories, and the varying interfaces of different publishers. Searching across platforms with distinct search functionalities consumes significant time and effort, particularly as cybersecurity norms and guidelines evolve rapidly, necessitating constant updates. Additionally, repositories often include documents from multiple domains beyond cybersecurity, which complicates the search for relevant resources.

The challenge of analyzing and implementing technical and legal documentation extends beyond cybersecurity to fields such as healthcare, engineering, and the social sciences. The specialized nature of these documents, coupled with sector-specific methodologies and regulations, adds another layer of complexity. Interpreting and applying the content across different organizational contexts requires a deep understanding of both technical and legal nuances. This complexity often leaves users unaware of their rights, legal protections, and obligations regarding personal data and digital services [6,7].

Emerging text ingestion tools, such as advanced NLP models, offer a solution to the challenges in collecting, processing, and analyzing documentation. By automating the abstraction and summarization of large volumes of text, these tools facilitate the quicker comprehension and application of complex documents [8,9]. Models like Generative Pre-Training Transformer (GPT) 3.5 [10] and GPT-4o [11] utilize tokenization, context-aware embeddings, and attention mechanisms to efficiently process and distill key information, enabling professionals to navigate and analyze documentation with greater speed and accuracy. Tokenization breaks text down into smaller units, while context-aware embeddings represent these units based on the surrounding context. Attention mechanisms help the model to focus on the most relevant parts of the text, improving the summarization accuracy. By identifying key themes and relationships, these models generate concise summaries that capture the essence of complex documents, making information more accessible and easier to apply.

The potential of NLP models has been demonstrated across various domains. Ghumade et al. [12] developed an NLP-based document classification system that outperformed other models available at the time. Cascella et al. [13] explored the benefits and challenges of using recent NLP models like ChatGPT 4 in healthcare, highlighting their usefulness in manual curation, interpretation, and knowledge discovery within the biomedical literature. Merchant et al. [14] created an NLP model for the summarization of legal documents, aiding lawyers in analyzing cases. This tool, developed before GPT's launch, received positive feedback from legal professionals. Ref. [15] proposed a solution combining GPT-3.5 Turbo and the Donut NLP model to classify civil construction specifications and extract data from images, ultimately creating a JSON-structured table of contents to simplify document navigation.

In the construction field, Saka et al. [16] identified documentation management as a key area where GPT can assist professionals. Savelka et al. [17] conducted a comprehensive analysis of GPT-4's ability to annotate legal texts, comparing its performance to that of law students. While GPT-4 showed acceptable results in this specialized task, improvements are needed, and the use of batch predictions was found to be more cost-effective, with a minor performance tradeoff. Liu et al. [18] addressed token limitations in NLP models

by implementing semantic clustering to reduce the document size, enabling the effective summarization of large texts within GPT's constraints. Alada et al. [19] demonstrated the innovative use of GPT for historical research by analyzing a 17th-century Ottoman text, proving that these technologies can provide valuable insights in various fields, including historical studies.

Thippeswamy et al. [20] presented the Streamlit application for text and PDF analysis, showing that LLMs improve the analysis accuracy and depth while emphasizing the need for advancements in LLM integration and privacy preservation. In [21], Vallabhaneni et al. introduce a system for automated document processing and advanced question answering that uses the Mistral 7B LLM for accurate and context-aware chatbot responses. Lin et al. [22] introduce "GPT4ESG", a model that combines various NLP architectures to enhance environment, society, and governance reporting analysis. A "semantic textual similarity analysis" of medical and biomedical texts was performed by Feng [23], where new LLM technologies were applied, and promising results were obtained. Ibrahim et al. [24] demonstrated how businesses can develop LLM-based chatbots capable of analyzing, summarizing, and extracting insights from textual data. The use of LLM models in analyzing historical legal documents was evaluated by Litaina et al. [25], where it was concluded that LLM models outperformed human experts in several aspects.

In Ref. [26], Bouzid et al. explored the use of the GPT-4o LLM model to enhance document representation for improved information retrieval, particularly in short documents with limited features. Mao et al. [27] evaluated the performance of nine LLMs in generating timeline summaries from construction delay documents. The findings highlight the potential of open-source LLMs for document construction analysis, emphasizing the importance of model selection based on accuracy, efficiency, and resource constraints. The implementation of LLM models to process semi-structured data in PDF files related to sports club activities was developed by [28]. The study concluded that this approach could reduce the analysis time by 90%. Wiest et al. [29] explored the use of LLM models to anonymize medical records, facilitating the construction of medical-related datasets that ensure patient privacy while remaining useful for medical researchers. They concluded that LLM models can indeed be a valuable tool for this purpose.

We have analyzed the research works described above in several dimensions. The comparison of these works and our repository is summarized in Table 1. The dimensions under evaluation were as follows.

- LLM Models indicates whether the referenced works implemented the most recent models available (contemporary to GPT-3.5 Turbo and later).
- Document Classification specifies whether the authors developed document classification functionalities in their work.
- Provides Insights indicates whether the proposed NLP-based system extracts relevant information from the documents.
- LLM-Based Repository applies to works where a documentation repository was built using LLM technologies.
- Extra Features applies to works that included additional functionalities, such as user interfaces, statistical analysis, document filters, and other enhancements integrated with LLM applications.

Table 1. A comparative table between the studies presented in the literature and our work.

Proposal	Latest LLM Models	Document Classification	Provides Insights	LLM Based Repository	Extra Features
Ghumade et al. [12]		x	x		
Cascella et al. [13]	x	x	x		
Merchant et al. [14]		x	x		
Feyisa et al. [15]	x	x	x	x	
Saka et al. [16]	x				
Savelka et al. [17]	x		x		
Liu et al. [18]	x		x		
Aladag et al. [19]	x		x		
Thippeswamy et al. [20]	x		x		x
Vallabhaneni et al. [21]		x	x		
Lin et al. [22]	x	x	x		
Feng et al. [23]	x		x		
Ibrahim et al. [24]	x	x	x		x
Litaina et al. [25]	x		x		
Bouzid et al. [26]	x		x		
Mao et al. [27]	x				
Merilehto et al. [28]	x	x	x		
Wiest et al. [29]	x	x		x	
Our repository	x	x	x	x	x

The repository presented in this paper distinguishes itself from previous research described above by focusing on the integration of NLP models into a system designed specifically for collecting, classifying, and analyzing cybersecurity documentation. It combines NLP capabilities with visualization tools and automated document collection to present cybersecurity information intuitively and efficiently using intuitive graphs and visual interfaces. The system was tested in the context of consolidating regulatory, legal, and technical cybersecurity legislation, along with technical documentation from Portuguese and EU repositories. It integrates search, automation, and NLP-driven analysis tools to provide users with timely information about the latest cybersecurity governance documentation. By addressing the current challenges in documentation management, the repository delivers a scalable platform that can extend beyond the cybersecurity field to various industries facing similar documentation complexities, namely health, engineering, the social sciences, and many more.

The contributions of this paper are summarized as follows:

1. An automated repository to efficiently manage large volumes of documents, which aggregates a wide set of cybersecurity-related documents from various sources in Portugal and the EU;
2. A set of user-friendly search and visualization tools to provide navigation and accessibility to documents;

3. A Git-Hub repository with the whole application and the tests developed so far, available at https://github.com/JoseMiguelFrade/Automated-Repository/tree/main (accessed on 1 March 2025).

The LLM approach used in the repository incorporates a real-world application of LLMs for the filtering and classification of documents from multiple sources and in various formats. The ultimate goal is to provide end users with the maximum amount of relevant information and insights while minimizing their effort, all through a single, AI-driven approach. The methodology adopted could pave the way for other AI-based solutions that simplify implementation processes while generating richer and more comprehensive outputs.

2. Background

This section aims to facilitate the understanding of this work by outlining the most relevant concepts and definitions.

2.1. Relevant Documents in Cybersecurity

Cyberlaw encompasses a comprehensive body of laws, decree laws, directives, and regulations developed to address challenges that arise within cyberspace. This legal domain covers a broad spectrum of issues, including cybersecurity, privacy, digital transactions, intellectual property, and access to digital resources.

The central objective of these legal frameworks is to protect individuals, organizations, and governments from cybercrime, data breaches, and other malicious online activities, while promoting the responsible and ethical use of digital assets and information.

Beyond legal instruments, technical standards and frameworks play a critical role by offering best practices and guidelines to help organizations to strengthen their cybersecurity and secure their digital infrastructure.

Cybersecurity regulations span various sectors, such as healthcare, defense, energy, telecommunications, and finance, reflecting the distinct vulnerabilities and needs specific to each industry.

Given the borderless nature of the internet, international treaties and regulations have been introduced to harmonize legal approaches, enhance cross-border collaboration, and improve collective efforts to mitigate cyber threats. As digital technologies continue to evolve, the significance of these legal frameworks in ensuring the governance and security of cyberspace grows, requiring continuous updates and refinements to address emerging risks and uphold the rights and safety of global internet users.

2.2. Issuer Organizations

In the context of Portugal and the European Union (EU), several key entities play a crucial role in issuing legal and technical documents and frameworks related to cyberspace. These entities include the following.

- **Legislation Issuers**
  - **Assembly of the Portuguese Republic** [30]: The legislative body responsible for creating and approving laws in Portugal,including those related to cybersecurity.
  - **Portuguese Government** [31]: The executive branch that implements and enforces laws, including policies and regulations concerning cybersecurity.
  - **European Council** [32]: An EU institution responsible for setting the general governance of the European Union, including broad cybersecurity strategies.
  - **European Commission** [33]: The executive branch of the EU that is responsible for proposing legislation, implementing decisions, and managing the day-to-day business of the EU, including cybersecurity regulations and directives.

  - **European Parliament** [34]: The directly elected legislative body of the EU that works with the European Council to adopt and amend proposed legislation.

- **Technical Norms and Framework Issuers**
  - **The Portuguese Cybersecurity Center (Centro Nacional de Cibersegurança—CNCS)** serves as Portugal's national authority for the promotion of cybersecurity and the strengthening of the country's cyber resilience. Its core functions include monitoring and responding to cyberthreats, raising public awareness, offering expert guidance, collaborating with international partners, and supporting regulatory compliance [35].
  - **The European Union Agency for Cybersecurity (ENISA)** is the cybersecurity authority for the EU, committed to enhancing and coordinating cybersecurity efforts across member states [36].
  - **The International Standards Organization (ISO)** plays a significant role in cybersecurity by developing and promoting international standards that ensure the security and resilience of information systems and networks [37]. The most relevant standards related to cybersecurity and information security belong to the ISO-27000 standards family [38].
  - **The Payment Card Industry Security Standards Council (PCI SSC)** is a global framework designed to ensure the security of credit and debit card transactions and protect cardholder data. The PCI DSS provides a comprehensive set of requirements for the enhancement of payment card security [39].

2.3. Legal and Technical Documents

A variety of legal documents are issued to govern cyberspace. In addition, numerous organizations publish technical documents designed to assist entities in enhancing their IT infrastructures. The most relevant types of legal and technical documents are summarized below.

2.3.1. Laws and Decree Laws

In Portugal, laws and decree laws hold equal authority, with the main difference being in their approval processes: laws are approved by the Assembly of the Republic, while decree laws are issued by the government. When a conflict arises between a law and a decree law on a particular issue, the choice of which to apply depends on two considerations: the more recent legislation or the one that is more specifically relevant to the issue in question [40].

2.3.2. Regulations

The purpose and significance of issued regulations differ depending on whether they are issued by individual countries, such as Portugal, or by supranational entities like the EU.

When issued by countries, regulations serve as documents primarily designed to facilitate the practical application of existing laws and decree laws. They are not intended to introduce new legislation; instead, they offer detailed guidance on how to apply existing laws in practical scenarios. In some cases, specific laws come into effect only after the publication of regulations clarifying the methods for their implementation [41].

If published by the EU, a regulation becomes a binding legislative act, and it must be fully applied throughout the EU [42]. So, in other words, EU regulations introduce new laws and legal procedures that should be fully implemented by all countries in the EU.

2.3.3. Directives

Directives are documents issued by the EU with the goal to provide guidelines for member countries to follow. The main difference between EU regulations and directives is that directives offer more flexibility to member states in implementing the required legislation, allowing them to adapt the procedures to their national contexts [42].

2.4. Technical Standards and Frameworks

Technical standards and frameworks, while not legal documents, are essential in this field. These documents are issued by official entities accredited to do so and establish rules, guidelines, or specifications for materials, products, processes, or services, such as the ISO standards [43].

2.5. Repositories

The resources related to cyberspace governance and regulation are scattered across various repositories of legal and technical documents. The main repositories that can be found online are as follows.

- StandICT [44]: This repository belongs to the European Standardization Observatory and contains many IT-related standards organized by different sectors.
- Cyber Policy Portal [45]: This is a United Nations (UN) portal that offers a comprehensive platform for access to a wide range of documents related to the regulation of cyberspace. The portal offers valuable insights into the institutions responsible for managing cyberspace in each country and includes detailed information on international agreements concerning cooperation in cyberspace.
- ENISA Repositories: ENISA offers repositories containing technical documents issued by EU member countries or by ENISA itself. Two repositories stand out: ENISA Publications [46], where ENISA publishes technical documents covering Europe's cyberspace, regulations, emerging threats, and studies related to these topics; and ENISA's National Cybersecurity Strategies map [47], which allows users to access documents pertaining to the cybersecurity strategies of all European countries.
- Octopus Cybercrime Community [48]: Managed by the Council of Europe, this portal hosts an extensive repository of cybercrime and handles digital evidence information, namely measures, policies, and legislation adopted by several countries world wide.
- OneTrust DataGuidance [49]: DataGuidance is a platform that provides an extensive range of information on cyberlaw, specifically focusing on data privacy legislation. It includes news, articles, and discussions conducted by experts in the field, addressing the major institutions, policies, and legislation that govern cyberspace and regulate data privacy.
- Eur-Lex [50]: The EUR-Lex repository is an online resource that provides access to EU regulations, directives, decisions, legislation, international agreements, and preparatory acts. The platform is designed to serve a wide range of users, including legal professionals, researchers, and other practitioners, by providing advanced search tools that facilitate the retrieval of documents.
- CNCS Observatory [51]: The Portuguese Cybersecurity Center periodically publishes reports and insights on the state of cybersecurity in Portugal. These reports cover a wide range of topics, including the general cybersecurity landscape in Portugal, the state of cybersecurity in specific sectors, and the economic impact of cybersecurity, as well as emerging threats and cyber conflicts. The CNCS also publishes technical norms aimed at strengthening the cybersecurity capabilities of national organizations. A prime example of this effort is the "Quadro Nacional de Referência para a Cibersegurança" [52] (National Cybersecurity Reference Framework). This framework serves

as a comprehensive guide for organizations to enhance their cybersecurity measures and resilience.

- Diário Da República [53]: Diário da República's portal is the official publication platform of the Portuguese Republic. The portal provides online access to all types of legal documents, including laws, decrees, resolutions, and regulatory norms issued by various branches of the government. The portal is divided into two series: one contains laws, decrees, and other acts of general legislative nature, while the other includes a wide range of other official documents, such as notices, declarations, and contracts.

2.6. The Necessity of Automated Tools for Document Analysis

The complexity of legal and technical documents in cyberspace is vast, addressing a wide range of intricate issues. This complexity highlights the need for automated tools to assist in analyzing and managing these documents. Such tools can greatly support professionals in ensuring accuracy, efficiency, and compliance, while streamlining interpretation and implementation processes. Automation helps to reduce the burden on specialists and improves overall document management in the cyberspace domain.

3. Overall Repository Architecture

The developed repository architecture is described in the following subsections. The deployment consists of diverse technologies and tools, each with unique capabilities, equipped with multiple functionalities and innovative features.

3.1. Repository Structure

The structure of the repository is depicted in Figure 1 and consists of three main components: the frontend, the backend, and the database.

All documents available in the repository are stored in a MongoDB database [54], described in Section 3.3. The backend interacts with the database to store new documents or perform operations on existing ones. The backend has three main software components.

- ChatGPT API: The utilization of the GPT Application Programming Interface (API) [55] (Section 3.2) allows the utilization of a large language model (LLM)-type AI model for the automation of various tasks related to document processing (Section 3.6).
- PDFCrawler: The PDF crawler is specialized in obtaining documents in PDF format from specified Uniform Resource Locators (URL) and is the focus of Section 3.4.
- Really Simple Syndication (RSS) Feed Consumer: This backend module is responsible for querying external repositories in search of news documents. The implementation of this functionality is described in Section 3.5.

The backend is responsible for receiving instructions from the frontend, processing requests, and returning results via its API, as shown in Figure 1. The API follows the Representational State Transfer (REST) standard and was developed using the Flask framework [56].

Regarding the frontend component, its primary role is to make available an intuitive and comprehensive interface. It enables users to manage existing documents and add new ones and provides key insights through the "Relation Graph" (detailed in Section 4.2) and the "Statistics" component (covered in Section 4.3). The frontend was developed using the "Vuetify 3" framework [57]. The Relation Graph utilizes the "vis-network" library [58], while the graphs in the Statistics component were created using the "chart.js" library [59].

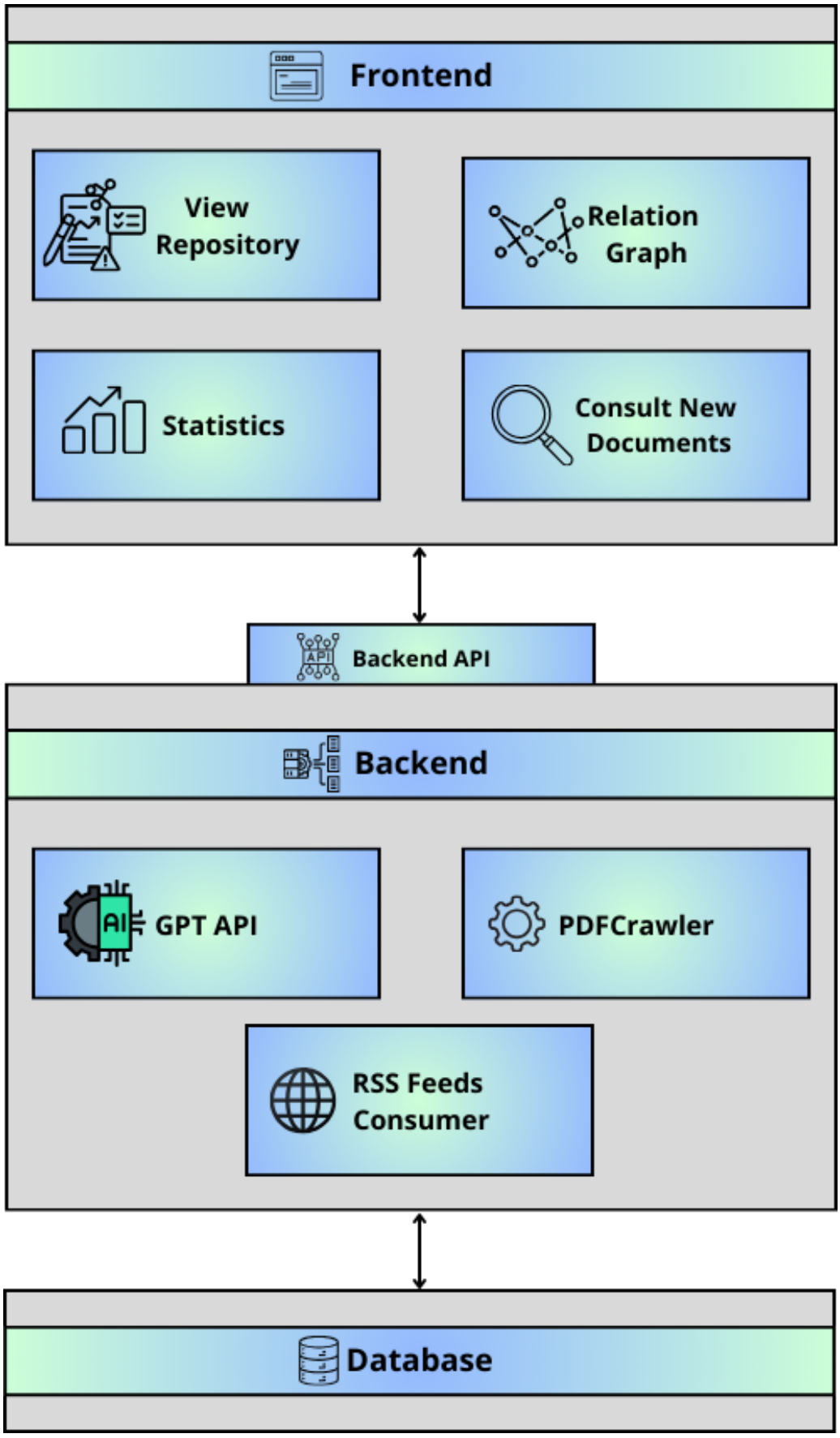

Figure 1. Automated repository implementation structure.

3.2. GPT API

ChatGPT [60], developed by OpenAI [61], is an LLM designed to generate human-like text. It belongs to the GPT series, which marks significant advancements in NLP technology. Based on the Transformer model, ChatGPT effectively handles sequential data like text.

The core mechanism behind ChatGPT's ability to generate coherent and contextually relevant responses lies in its two-stage training process: pre-training and fine-tuning. During pre-training, the model is exposed to a vast corpus of text data and learns to predict the next word in a sentence given the words that precede it. Fine-tuning then refines the model's responses by training it on smaller, task-specific datasets, improving its performance in specialized tasks like conversation, technical explanations, or creative writing. Several versions of ChatGPT are available today [62], including GPT-3.5, GPT-4, their "Turbo" variants, GPT-4o, and GPT-4o Mini.

For the automated repository, the GPT API enabled the integration of the GPT models' capabilities. Figure 2 shows an example API call to GPT-4o.

```
import openai
client = openai.OpenAI()
response = client.chat.completions.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": f"Analyze the following document extract:
        ↪ '{pdf_text}'. Is it related to IT/cybersecurity/data privacy/AI?"}
    ],
    temperature=0.4,
    top_p=1,
    max_tokens=250
)
return response.choices[0].message.content
```

Figure 2. Example of a call to GPT-4o API.

GPT API calls allow for the specification of several parameters, and the most important include the following.

- Model: Specifies which GPT model processes the request. Available models include "gpt-4o", "gpt-4-0125-preview" (GPT-4 Turbo), and "gpt-3.5-turbo-1106" (GPT-3.5 Turbo). Model versions and designations are regularly updated. This parameter is mandatory.
- Messages: Contains the prompts for the model. A message has two parts: "role" (context for the model) and "content" (the actual prompt). The first message should be in the System context ("role": "system") to guide the model to act like an assistant. Subsequent messages typically have the User context. This parameter is mandatory.
- Temperature: A value from 0 to 2 that influences the model's responses. Lower values make the model more factual and precise, while higher values make it more creative. This parameter is optional.
- Top_p: Similar to "Temperature", it controls the model's behavior by adjusting the probability distribution of tokens. It is recommended to adjust "Temperature" or "Top_p", but not both. The default value is 1.0. This parameter is optional.
- Max_tokens: Specifies the maximum number of tokens that the model can output. Tokens are basic units of text (words, parts of words, punctuation). This parameter helps to manage costs and, while not mandatory, is highly recommended.

When discussing the GPT API, pricing is a crucial factor. Pricing is based on the number of input tokens and output tokens. Although input tokens can be controlled by the request length, the number of output tokens is managed by setting the "Max_tokens" parameter. Costs vary by model, but input tokens are generally cheaper. Contrary to expectations, stronger models are not always more expensive. GPT-4o is cheaper than GPT-4 and GPT-4 Turbo, and GPT-4o mini is currently the most affordable model, offering a lower cost per token than GPT-3.5 Turbo [63].

3.3. MongoDB

MongoDB [54] is an open-source, non-structured query language database management system that uses a document-oriented approach to store data in flexible Javascript Object Notation (JSON) format documents. It is designed for scalability and the handling of diverse data types. For the storage of large files, MongoDB utilizes GridFS, a specification for the storage and retrieval of files that exceed the Binary Javascript Object Notation (BSON) document size limit of 16 MB. GridFS divides files into smaller chunks and stores them as separate documents, enabling efficient file storage and access within MongoDB. MongoDB's GridFS has been particularly useful in building the repository, as it is the technology currently used to store all documents available in the repository.

3.4. PDFCrawler

A primary objective in building the repository was to automate the collection of new documents. We adapted an already existing PDF crawler [64], which allows users to input the URLs of legal and technical repositories and retrieve all PDF files containing specific keywords from a configurable keyword list.

PDFCrawler employs a headless web browser (GECKO) to render websites and simulate clicks on all clickable elements. Users can set the search depth, with a minimum value of 1 indicating that only the current page will be crawled. For higher depth values, users can choose between two modes: "Crawl in Depth" or "Crawl in All Directions". The "Crawl in Depth" option restricts the crawler to sub-pages of the initial URL, following the pattern "/example/...". On the other hand, "Crawl in All Directions" allows the crawler to follow all URLs detected on the initial page, even if they do not lead to sub-pages of the initial URL. This provides a more comprehensive analysis but increases the task complexity. Higher depth values (3 or more) significantly increase the time and resources required for crawling.

All crawled documents are stored in local folders based on the domain from which they were retrieved. PDFCrawler handles the creation and management of these folders, generating a new one for each unique domain that it processes.

3.5. RSS Feed Consumer

Really Simple Syndication (RSS) is a web feed format used to publish frequently updated content like blog posts and news articles in a standardized XML format. The option of querying cybersecurity documents using an RSS feed was available in two repositories used in this project, DRE [53] and EurLex [50], each with unique specifications.

The RSS Feed Consumer helps users to stay informed about the latest cybersecurity documents relevant to both the user and the repository. When a user selects documents from the RSS Feed Consumer, the document links are sent to PDFCrawler, which retrieves the files for inclusion in the repository, seamlessly continuing the document addition process.

3.6. Automated Collection and Classification

To simplify document collection from multiple repositories, users can use PDFCrawler, as described in Section 3.4. Additionally, the RSS Feed Consumer, accessible through the "New Documents Page" (Section 4.4), keeps users updated on new documents relevant to the repository. After collection, the gathered PDF files are analyzed by GPT through the GPT API.

For each PDF, the initial step involves extracting the first 850 text tokens and sending this excerpt to the ChatGPT. Sending only excerpts helps to manage the costs and minimizes potential errors with larger documents, as models have token limits (128K tokens for GPT-4o and GPT-4 and 16K tokens for GPT-3.5 Turbo). We have conducted several tests to determine the optimal number of tokens to extract from each PDF. These tests involved gathering a diverse set of documents and then extracting a specific set of information from all PDFs using a single, identical query. The only variable across the operations was the number of tokens extracted from the documents and sent to the model.

It was observed that, after approximately 600 context tokens, the results were of good quality. The main issue with shorter context lengths was that the token count was often insufficient to provide the necessary information about the document to the model. By gradually increasing the number of tokens, it was found that the response quality did not vary significantly after reaching 750–800 tokens. The quality variation was even smaller for higher token values, such as 900 and 1000 tokens. This led to the conclusion that 850 tokens provide sufficient context for quality responses while avoiding an unnecessary amount of tokens being analyzed by the model each time.

OpenAI's Tiktoken library [65] was used to measure the number of tokens in text samples, ensuring consistency with the token counting methods of GPT models. Next, the 850 tokens are sent to a GPT model for analysis. To optimize the costs, 60% of the PDF files are analyzed using GPT-4o and 40% using GPT-4o mini. Users can adjust the number of PDFs analyzed and the frequency of analysis by each model. Figure 3 displays the repository's "Update Page" for these settings.

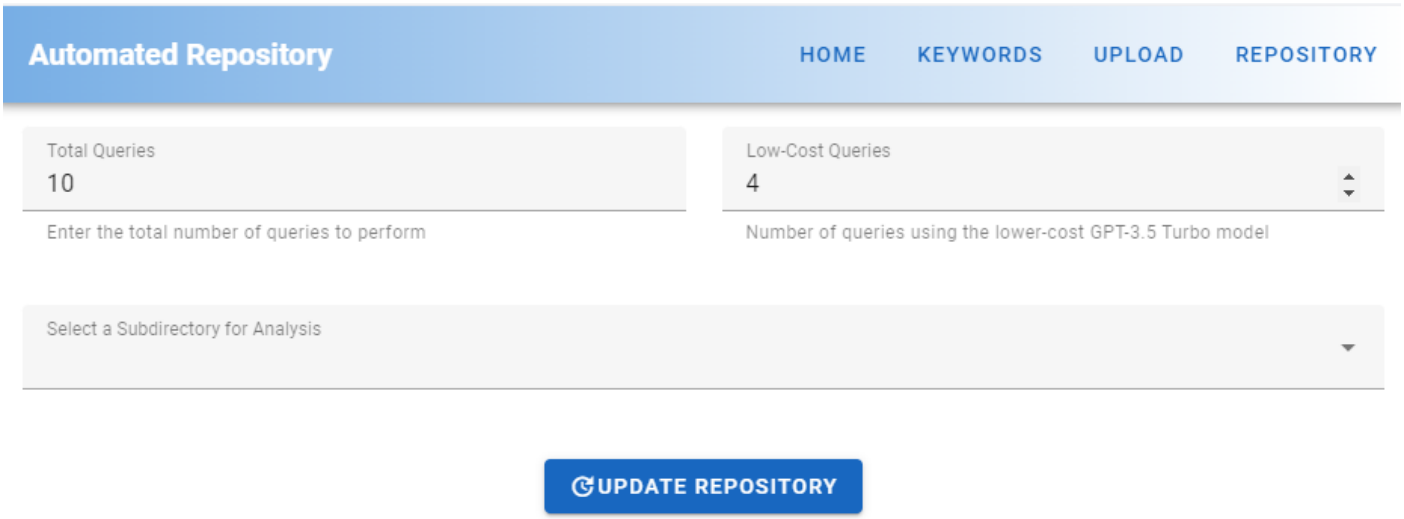

Figure 3. Update page.

For both models, the request message follows the format shown in Figure 4. For each PDF document, two prompts are sent to GPT. The first prompt instructs the model to act as an assistant, formatting responses as

```
"field:<field_value>"
```

While the GPT API offers a JSON response format option, it was observed that the use of this option caused issues, especially when the responses exceeded the *max_tokens* limit, leading to incomplete or invalid JSON outputs. Additionally, using the JSON output format consumed more tokens compared to our custom format, making the latter more efficient.

```
messages=[
{"role": "system", "content": "You are a helpful assistant. Always respond
in the format 'field:<field_value>'."},
{"role": "user", "content": f"Analyze the following document extract:
'{pdf_text}'. First, determine if it is related to
IT/cybersecurity/data privacy/AI.
If it is not related, respond with 'is_related:<no>'.
If it is related, provide structured information with the following format
(if no related_docs, the~value for related_docs is <none>)
(use English to write the Abstract and Type):
'is_related:<yes>#issuer:<issuer_name>
#origin:<origin>#type:<Norm/Law/Regulation/Treaty/...>
#subject:<Privacy/Governance/Security/...>#date:<date in dd/mm/yyyy format>
#area:<Finance/Healthcare/General/Energy/...>#title:<document_title>
#Related_Docs
|doc2|doc3#abstract:<brief_summary (95 tokens max)>'."}
],
```

Figure 4. Messages sent to GPT models.

The second prompt is tailored to the user's context, asking the model to classify the PDF based on the document excerpt: "Analyze the following document extract: 'pdf_text'". The model is first asked if the document pertains to topics related to cybersecurity, IT, data privacy, or AI. If the answer is "no", the model stops further analysis, conserving output tokens. If the answer is "yes", the model proceeds to provide the following information about the document.

- Title: Official document designation (e.g., Regulation (EU) 2016/679).
- Date: Date of first publication.
- Origin: Document origin (country, institution, etc.).
- Issuer: Entity that published the document (e.g., European Parliament).
- Subject: Document's focus (e.g., security, data privacy, governance).
- Area: Sector targeted by the document (e.g., healthcare, justice).
- Related Documents: Directly related documents, if any.
- Abstract: Brief summary (maximum of 95 tokens) of the document.

The retrieved information and corresponding PDF files are stored in the database along with the the file hash and the date that the file was uploaded to the repository. A simplification of the document collection and classification processes is presented in Figure 5.

The details of each document are stored in a unique MongoDB collection, and the PDF file is stored with the help of "fs.chunks" and "fs.files" collections. The "pdf_file_id" value links the document to its PDF file. An example of a document stored in MongoDB's "Documents" collection can be observed in Figure 6.

The process described in this section is represented in Figure 7. It starts with the user creating a keyword list for initial filtering. The user can then either crawl an online repository of their choice or consult recent documents identified by the RSS Feed Consumer. If the user chooses to consult the RSS Feed Consumer, the diagram shows two scenarios: either no relevant documents are found, requiring the user to restart the process, or relevant

documents are selected and sent to the crawler. If the user opts to crawl specific repositories, they input the repository links into the crawler as illustrated.

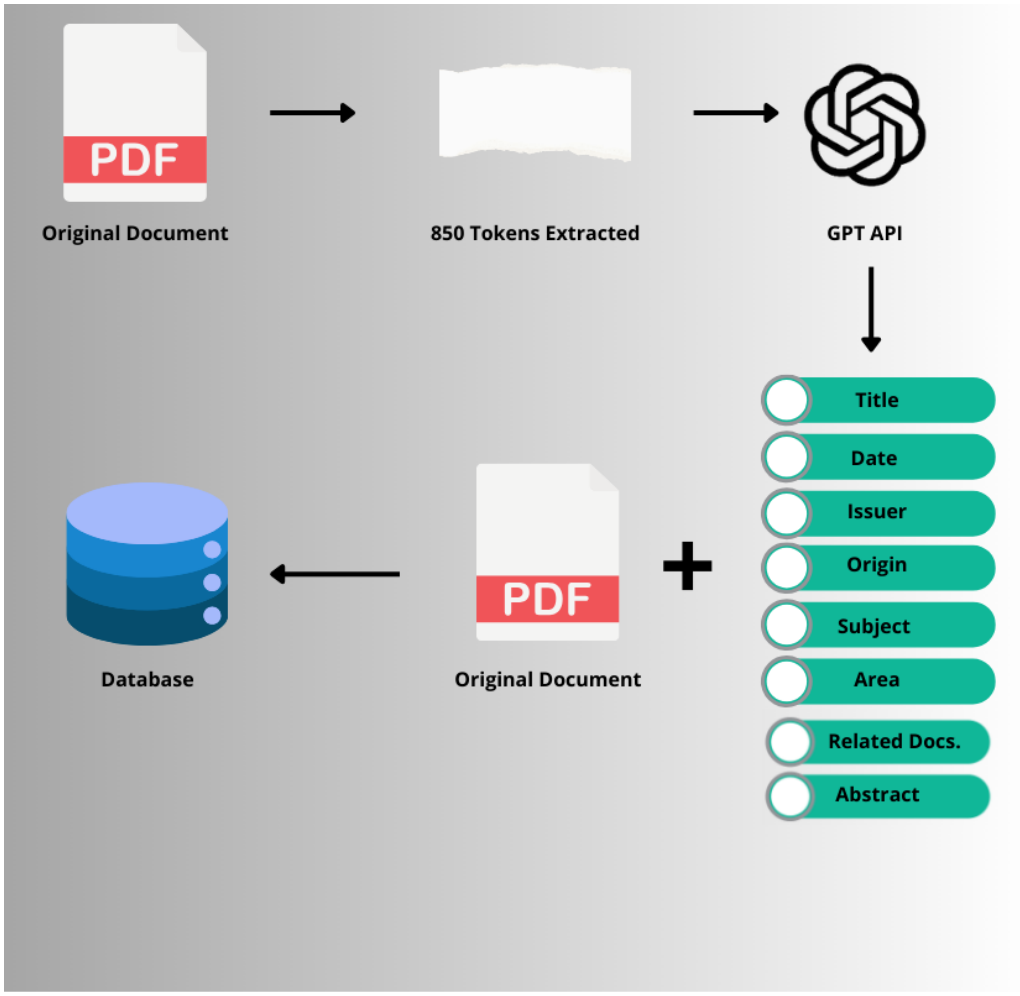

Figure 5. Simplification of document collection and classification processes.

After documents are collected and filtered using the keyword list, the application generates a hash for each document and checks for duplicates. Duplicates are excluded, while unique documents proceed, with their first 850 tokens sent to the GPT API for evaluation. If the document is related to cybersecurity, GPT categorizes it, and the information is stored alongside the document's hash, upload date, and PDF file in the database.

For post-analysis operations, the repository includes four folders for the storage of copies of crawled documents: "Accepted", "Rejected", "Manually Deleted", and "Duplicated". Documents that were rejected during keyword filtering or GPT filtering are stored in the "Rejected" folder. However, if they were rejected during the hash comparison, they are stored in the "Duplicated" folder. Accepted documents are classified by GPT and stored in the repository, with one document's copy being stored in the "Accepted" folder. If, later, the document is deleted from the repository, the document's copy is moved to from the "Accepted" to the "Manually Deleted" folder. This system allows for analysis of the acceptance and rejection rates, as well as tracking how many accepted documents are later deemed irrelevant and deleted.

```
{
"_id": {
"$oid": "65de3ca7fcc0a2399f08395c"
},
"is_related": "yes",
"issuer": "European Parliament and Council of the European Union",
"origin": "European Union",
"type": "Regulation",
"subject": "Cybersecurity",
"date": "17/04/2019",
"area": "Cybersecurity",
"title": "REGULATION (EU) 2019/881",
"related_docs": [
"Regulation (EU) 526/2013",
"Regulation (EU) 2016/679",
"Directive (EU) 2016/1148",
"Directive (EU) 2018/1972"
],
"abstract": "The REGULATION (EU) 2019/881 outlines a comprehensive framework for the
↪ European Union's cybersecurity through the establishment of the European Union Agency
↪ for Cybersecurity (ENISA). It covers the agency's mandate, objectives,
↪ and˜organizational structure, including the management and executive boards. Key
↪ areas addressed include policy development, capacity-building, operational
↪ cooperation, market certification, awareness-raising, research, and˜international
↪ collaboration. The˜regulation also details the establishment and implementation of
↪ European cybersecurity certification schemes, protection of sensitive information,
↪ and˜ENISA's budget and staffing.",
"pdf_file_id": "65de3ca7fcc0a2399f083957",
"pdf_hash": "774ca4e4995defd57f6235a0c3b21fff",
"upload_date": "19:48:55 27/02/2024"
}
```


Figure 6. Example of a document stored in MongoDB's documents collection.

3.7. Implementation Scenarios

On the GitHub page (https://github.com/JoseMiguelFrade/Automated-Repository/tree/main (accessed on 1 March 2025)), there is a list of requirements to run the application, a set of recommendations, two installers (one for Windows and another for Linux), and two launchers (one for Windows and another for Linux). The capabilities of the repository were fully explored on a Windows 11 machine with 16 GB of RAM and an 8-core processor. The results obtained with this setup are discussed in Section 5. The repository was also tested on a Windows 11 machine with 4 GB of RAM and a 2-core processor and on a Ubuntu 24.04 desktop.

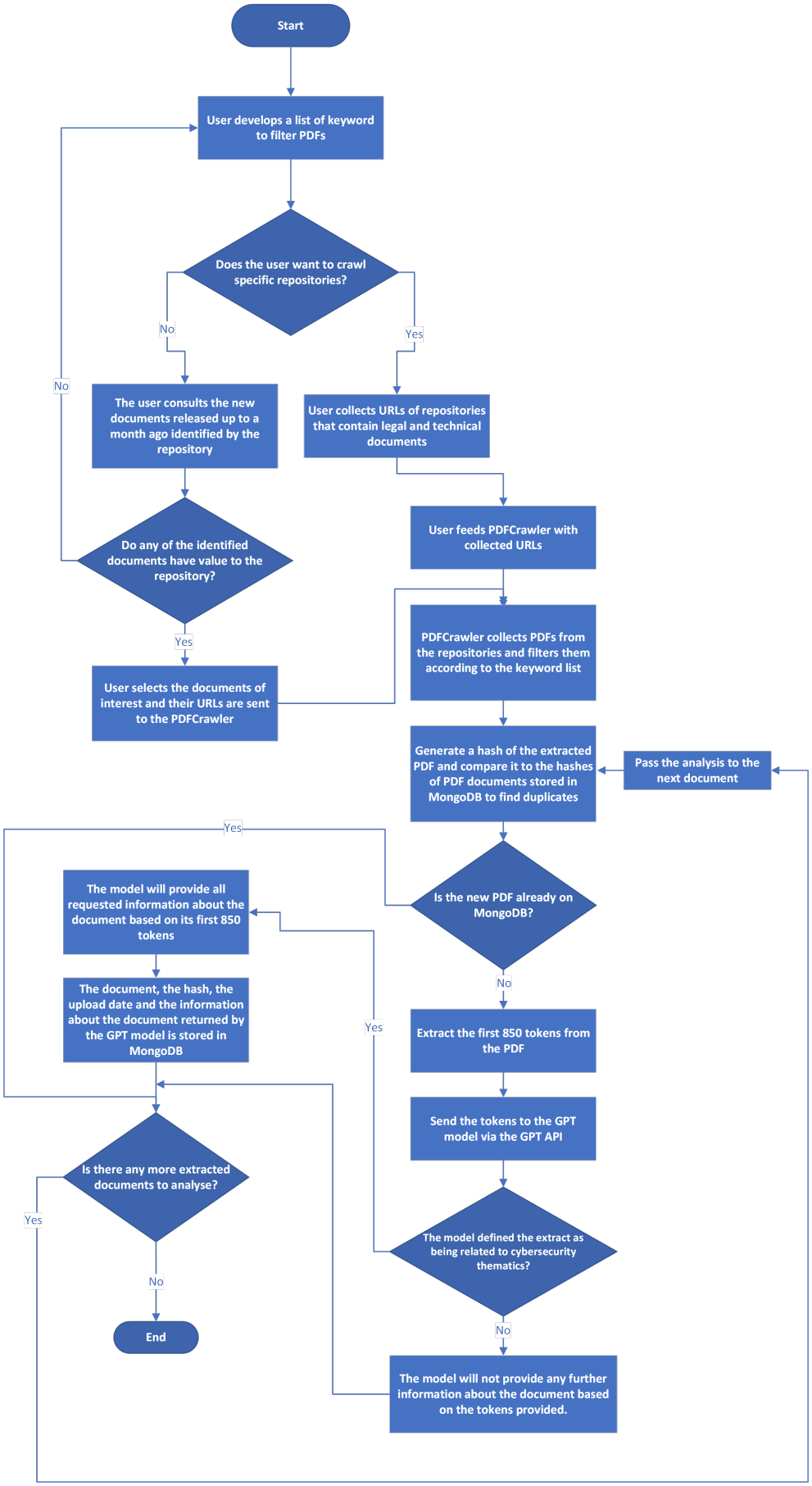

Figure 7. Flowchart for the process of adding documents to the automated repository.

4. Result Analysis

This section details the main functionalities available in the repository and the main corresponding results obtained in terms of downloaded and categorized documents. The home page shows all documents contained in the repository, represented as unique cards. It has pagination, a search bar, and filtering tools (Figure 8).

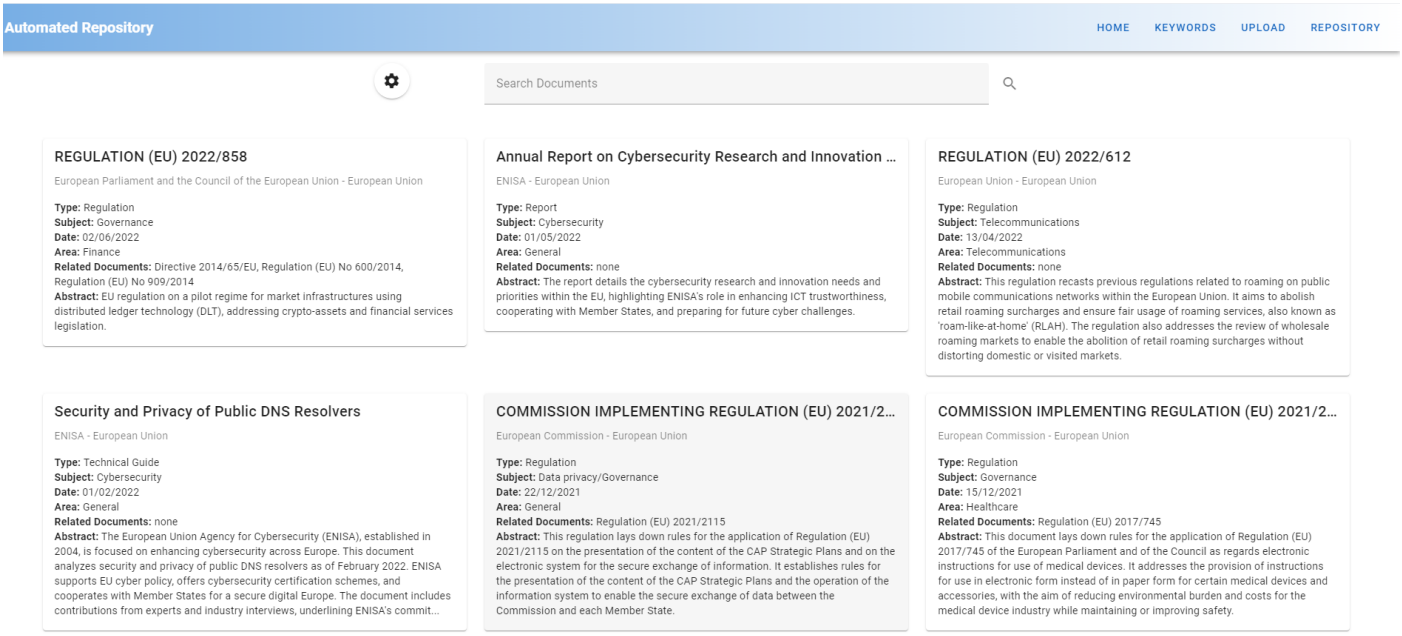

Figure 8. Repository's main page.

The interface fetches documents via a `HTTP GET` request to the backend's API endpoint "/get-documents". Upon receiving the request, the backend queries the database and returns all documents to the frontend in a JSON list of objects, where each object represents a document and its related information. By clicking on a specific card/document, users are redirected to a "Details Page" that presents all the information related to that document, including also the fields that were generated by GPT.

In the "Details Page", users have the ability to perform a variety of operations on the document that they are currently viewing. Users can manually edit all fields of the document, ask GPT to regenerate a specific field (Section 4.1), download the original document, and delete the document from the repository.

4.1. Regenerate Document Fields

A core functionality of the repository is the "Regenerate Page", where users can ask GPT to reanalyze any field of a document and set a specific "creativity level" for the operation. As shown in Figure 9, the field values of the current document are displayed. These values are obtained by a `GET` request to the "/get-document/documentId endpoint". The page also includes a slider that allows users to adjust the creativity level for the reanalysis.

Figure 9. Regenerate page.

After selecting the field to be regenerated, a POST request will be sent to the backend through the "/regenerateDoc" endpoint. The request structure follows the example shown in Figure 10. As can be observed, three values are sent in the request.

- documentID: This is the ID that represents the document whose field is going to be regenerated.
- field: This is a value that represents the field that will be regenerated. This information will be used to select the correct prompt that will be sent to GPT to perform this operation.
- temperature: This is the creativity level, also known as the temperature, as explained in Section 3.2, at which GPT operates while generating its response. This level ranges from 0 (more technical) to 2 (more creative).

```
{
  "documentId": "65e6fa22803d02a35e8e21fc",
  "field": "abstract",
  "temperature": 0.8
}
```

Figure 10. Regenerate document POST request.

After receiving the `POST` request, the backend will prepare an excerpt comprising the first 850 tokens from the document. During this operation, the value of "documentID" from the POST request is utilized. Alongside this excerpt, a prompt will be crafted to instruct GPT to analyze the excerpt and address the user's query about the selected field. The information about the selected field, as contained in the `POST` request, will be essential for this task. Finally, the document excerpt and the request will be sent to GPT via its API by the backend. All requests will be analyzed by the GPT-4o model.

The user has always the option to revert to the original value of the field that was first displayed when they opened the page. This can be achieved by selecting the back arrow below the regenerated field, as shown in Figure 9.

4.2. Relations Graph

In addition to displaying and describing the documents, the repository includes the functionality to showcase the relationships between them. This is achieved through a network graph, which is illustrated in Figure 11. At the top of the Figure 11 there is a high-level view of the graph, while at the bottom, a detailed view is shown, which can be obtained by simply zooming in on the graph.

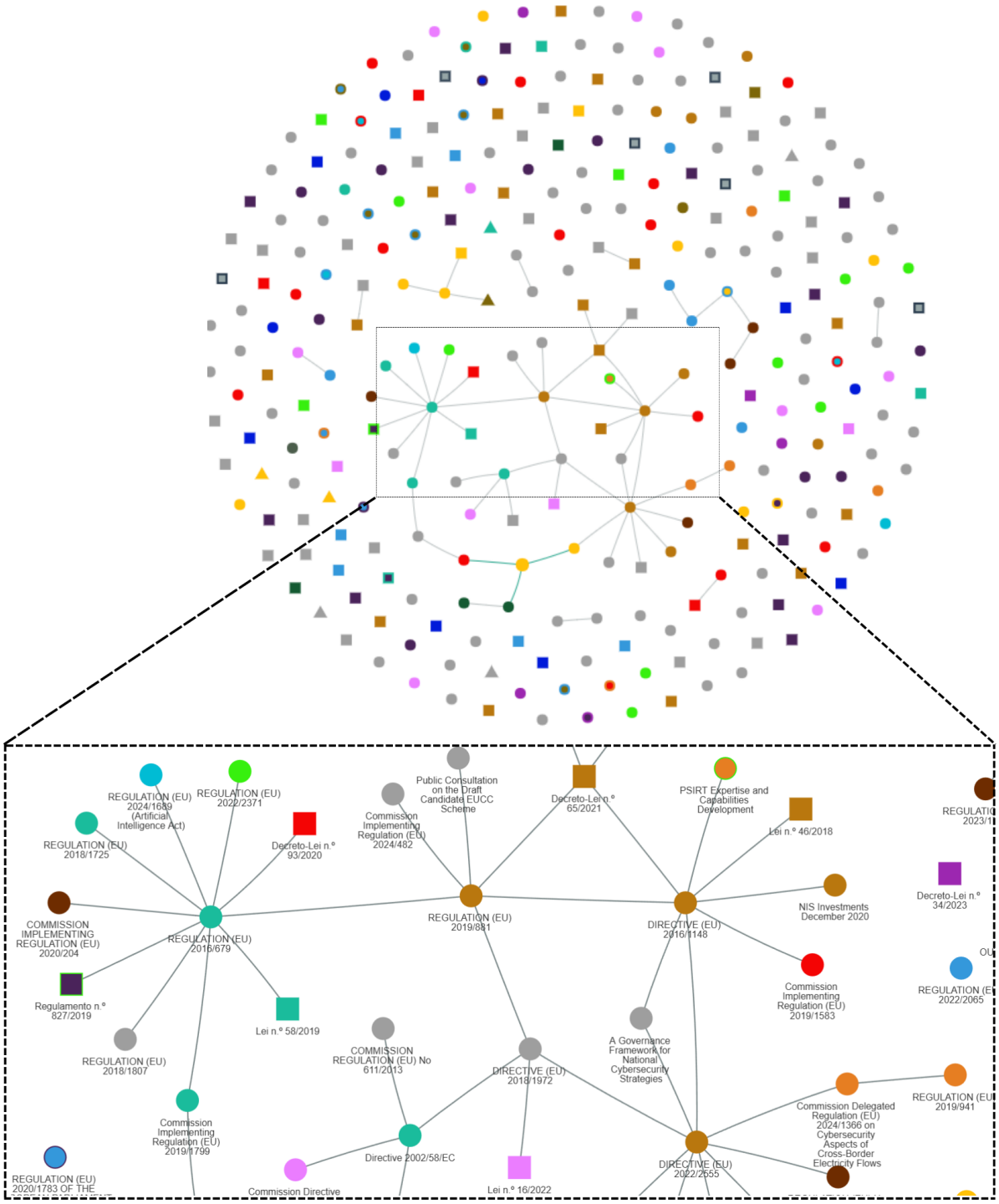

Figure 11. Relation graph.

Each node in the graph represents a single document, and each connection between two nodes represents a relationship between the two documents. The relationships depicted in the graph are established based on the "Related Documents" field of each document. For example, if a document lists "REGULATION (EU) 2016/679" in this field, a relationship will be illustrated in the graph between that document and the document "REGULATION (EU) 2016/679". This connection is represented even if "REGULATION (EU) 2016/679" does not reference the first document in return. However, if a document references a related document that is not present in the repository, this particular relationship will not be depicted in the graph.

Figure 12 represents a segment of the relations graph, where it is possible to observe that, in addition to showing relationships, each node possesses additional characteristics. Below each node is the title of the document that it represents. Furthermore, it is evident that the nodes differ in color and shape. The colors of the nodes indicate the areas of the documents that they represent. Each area is associated with a specific color, namely cybersecurity, artificial intelligence, digital rights, or healthcare.

Figure 12 also reveals that some nodes have a colored border. This border color denotes the secondary area of the document that originated the node. This occurs because every document may be associated with a maximum of two distinct areas.

The shapes of the nodes signify the origin of the document. In Figure 12, two different shapes can be observed: squares and circles. Squares indicate that the documents originated from Portugal, while circles signify that they come from a European Union institution.

Users can also click on each node. If a click is detected, the user interface will redirect the user to the "Details Page" of the document with the ID corresponding to the ID of the document represented by the node. Finally, users can search for a specific node on the graph by entering the desired document's title in the search bar. If a document with the corresponding title exists, the graph will zoom in on this specific node, and the node will be highlighted.

The relations graph provides valuable insights to users, as it shows how cybersecurity documents relate to each other. It also shows that there are some documents that are central in cybersecurity documentation, as they are related to many other documents and their overall number of relations is far superior when compared to many other documents.

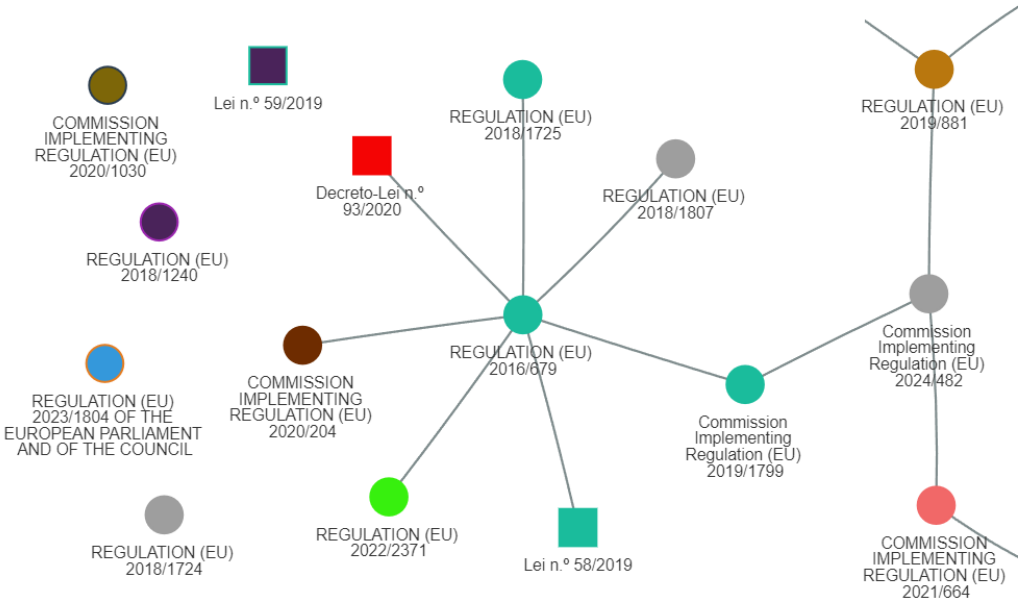

Figure 12. Relations graph detail.

4.3. Statistics

The repository has a "Statistics Page" containing a series of data visualizations that provide users with insights into the repository's contents. The visualizations include the following:

- Number of documents grouped by area, from all or different issuers—Figure 13;
- Number of documents issued over time and grouped by different origins—Figure 14;
- Cumulative and monthly count of documents present in the repository—Figure 15;
- Number of documents grouped by type (law, decree law, report, etc.)—Figure 16;
- Number of documents issued per year and per area—Figure 17.

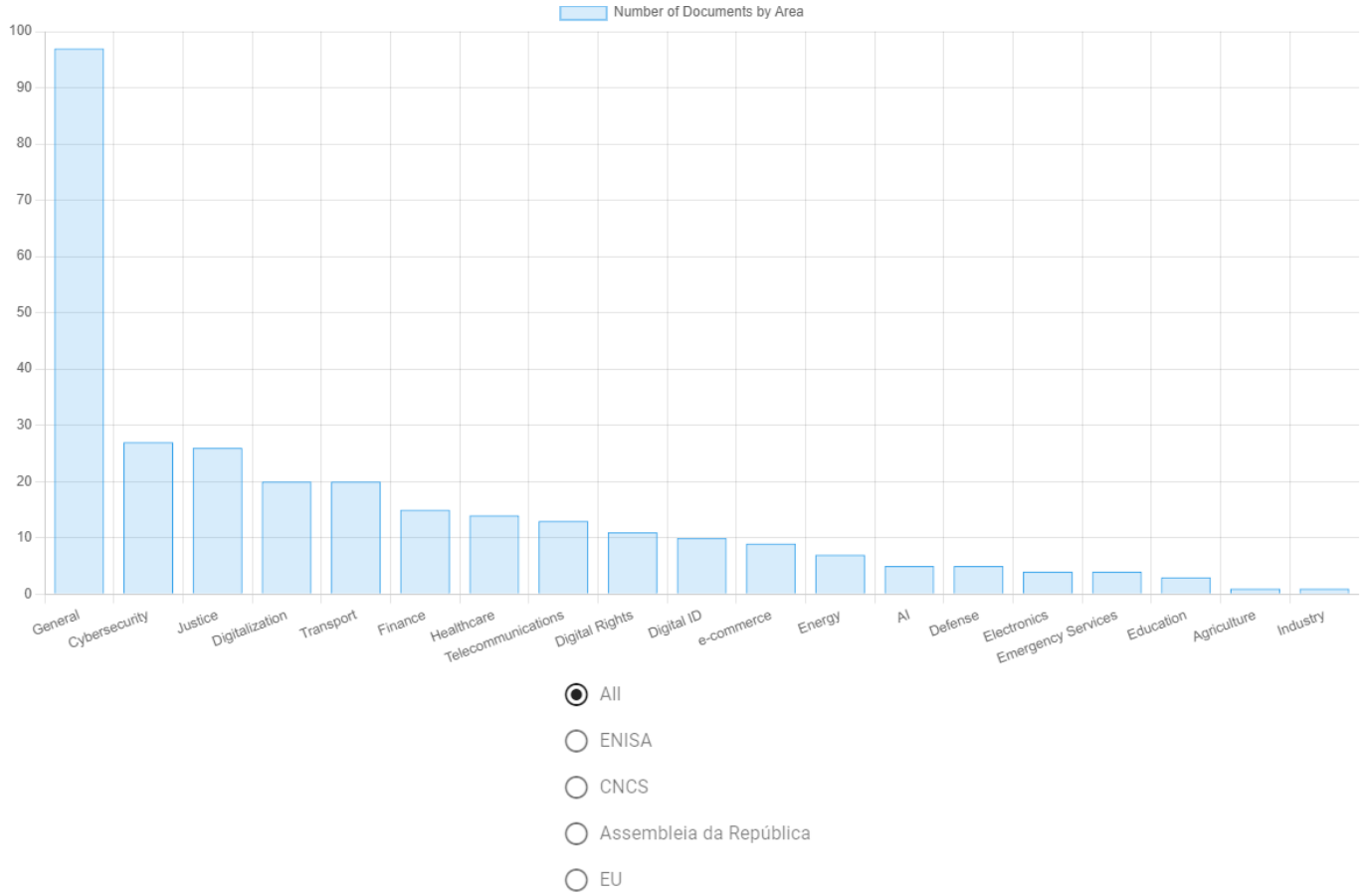

Figure 13. Number of documents grouped by area.

Users are able to gain a better understanding of the repository's contents and better comprehend the evolution of cybersecurity-related documents over time and their distribution across the various areas. This type of information is particularly useful in drawing insights about the evolution of these types of documentation and the areas that have been most impacted by new regulations and legislation.

4.4. New Documents Page

Users can provide URLs for repositories or documents relevant to the repository. Additionally, the "New Documents Page", which is represented in Figure 18, can notify users of the latest documents that might be of interest and can be added to the repository.

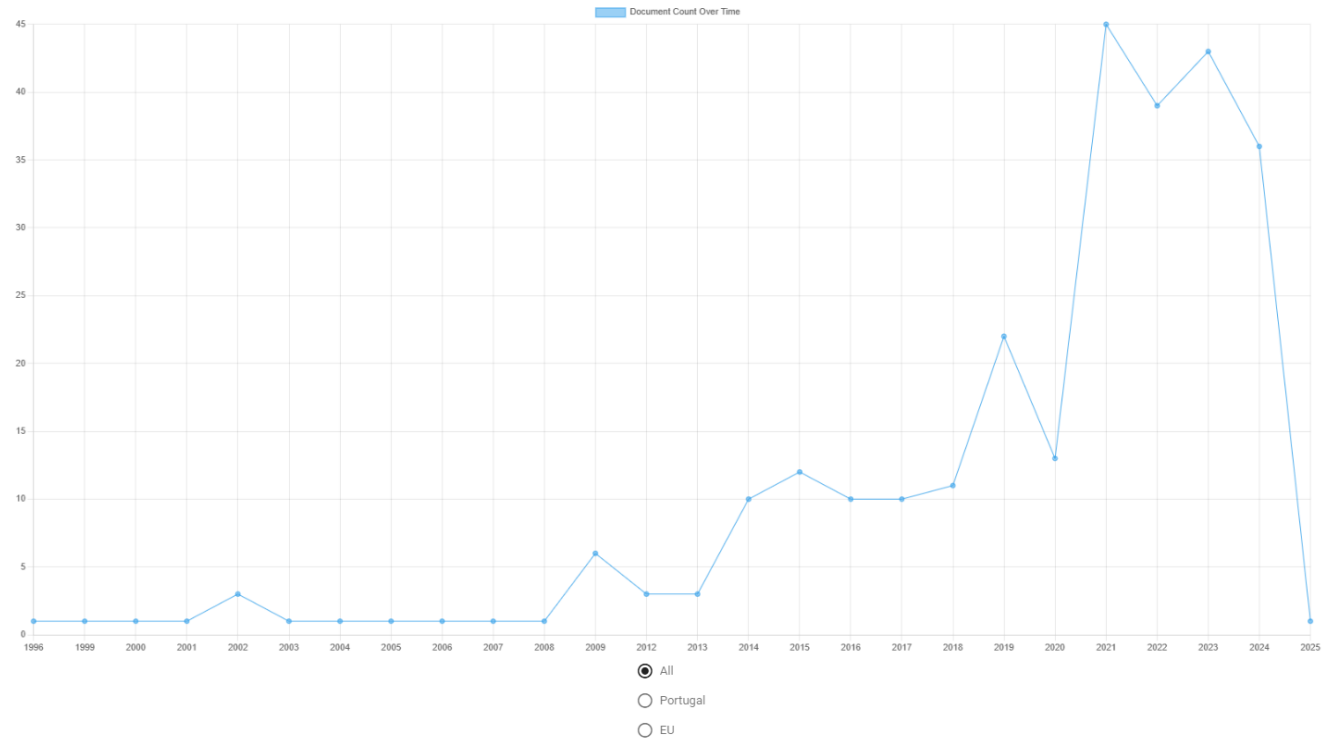

Figure 14. Documents issued over time and grouped by origin.

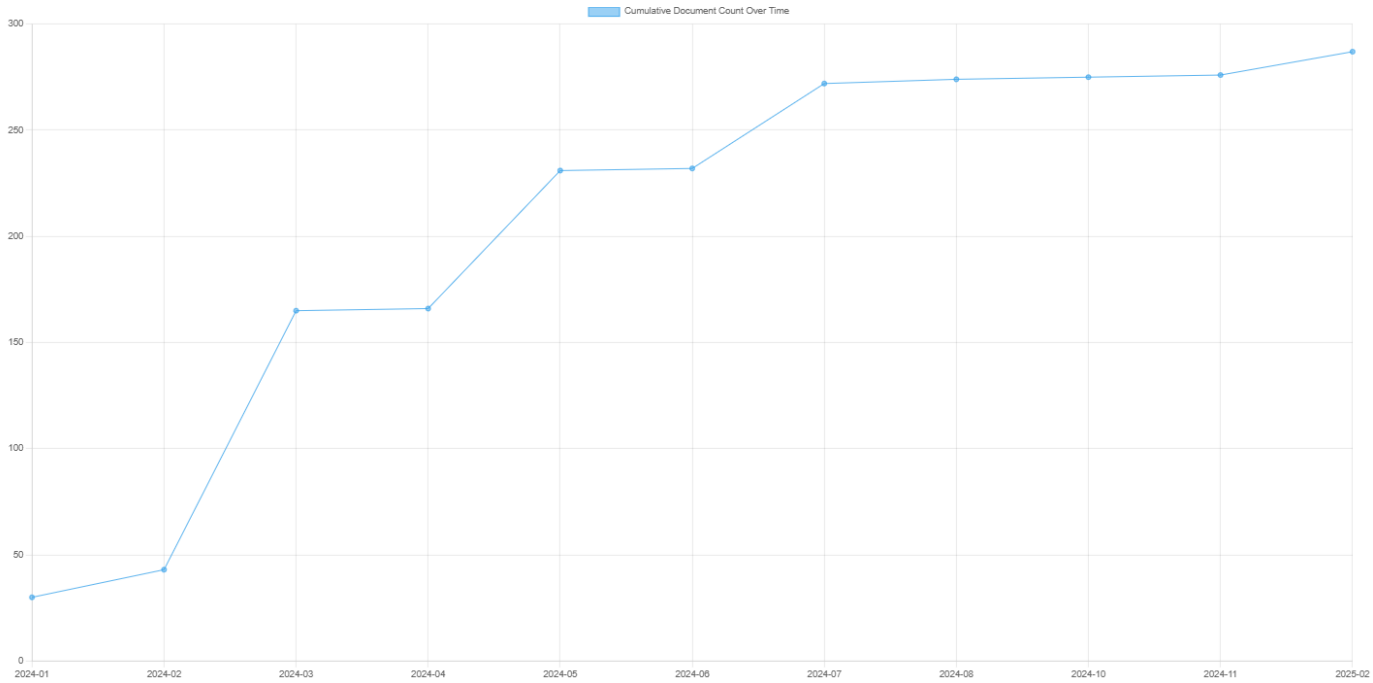

Figure 15. Cumulative count of documents present in the repository.

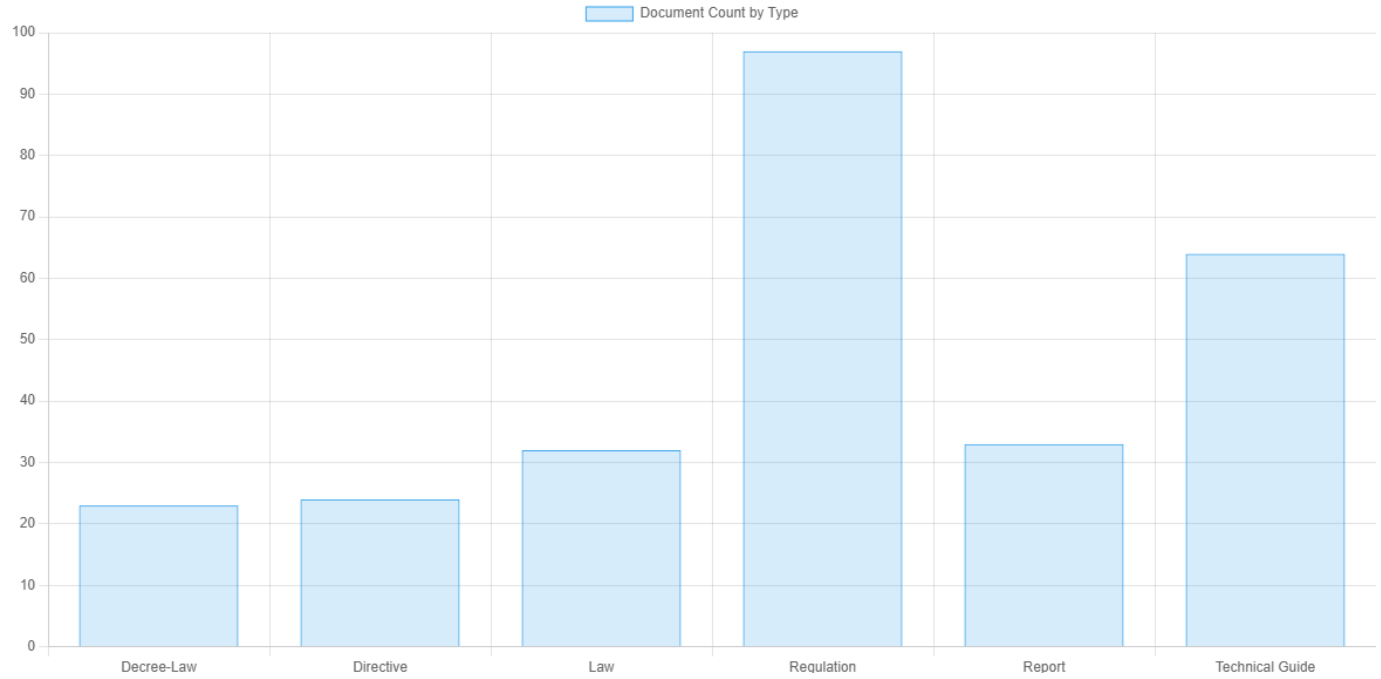

Figure 16. Number of documents by type.

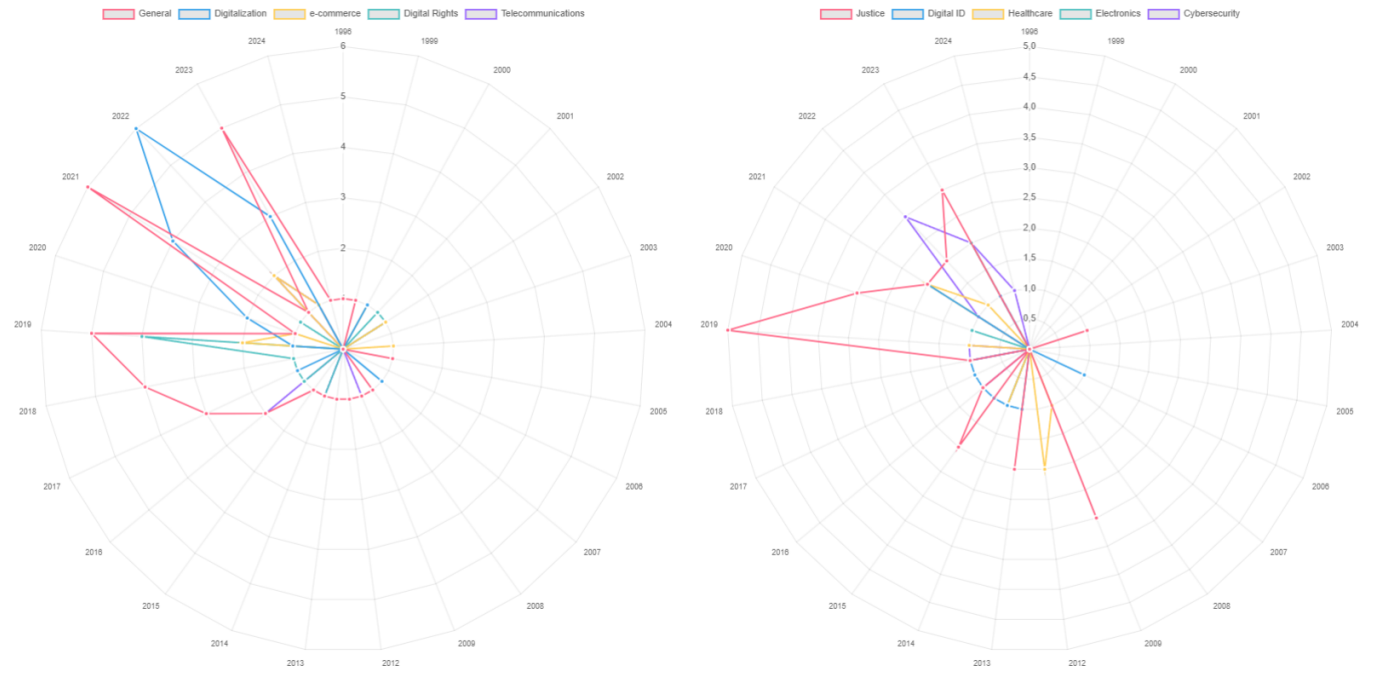

Figure 17. Number of documents issued by year and by area.

Using the "New Documents Page", users may select individual documents or groups of documents, each represented by a card, and choose the "Send to Crawler" option. This action sends the selected documents to the crawler (PDFCrawler) to fetch and store them for future analysis by the GPT API. All information about new documents (up to one month old) on the page is retrieved using RSS Feeds from EurLex and `dre.tretas.org`, as explained in Section 3.5. Each card displays the document's title, such as Law 54/2023 of September 4th (or Lei 54/2023, de 4 de Setembro), date, and a short description, helping users to easily determine the relevance of the document to the repository. (The reason the

document's information displayed in the cards is in Portuguese is that the data was directly extracted from the DRE repository, where all information is in Portuguese).

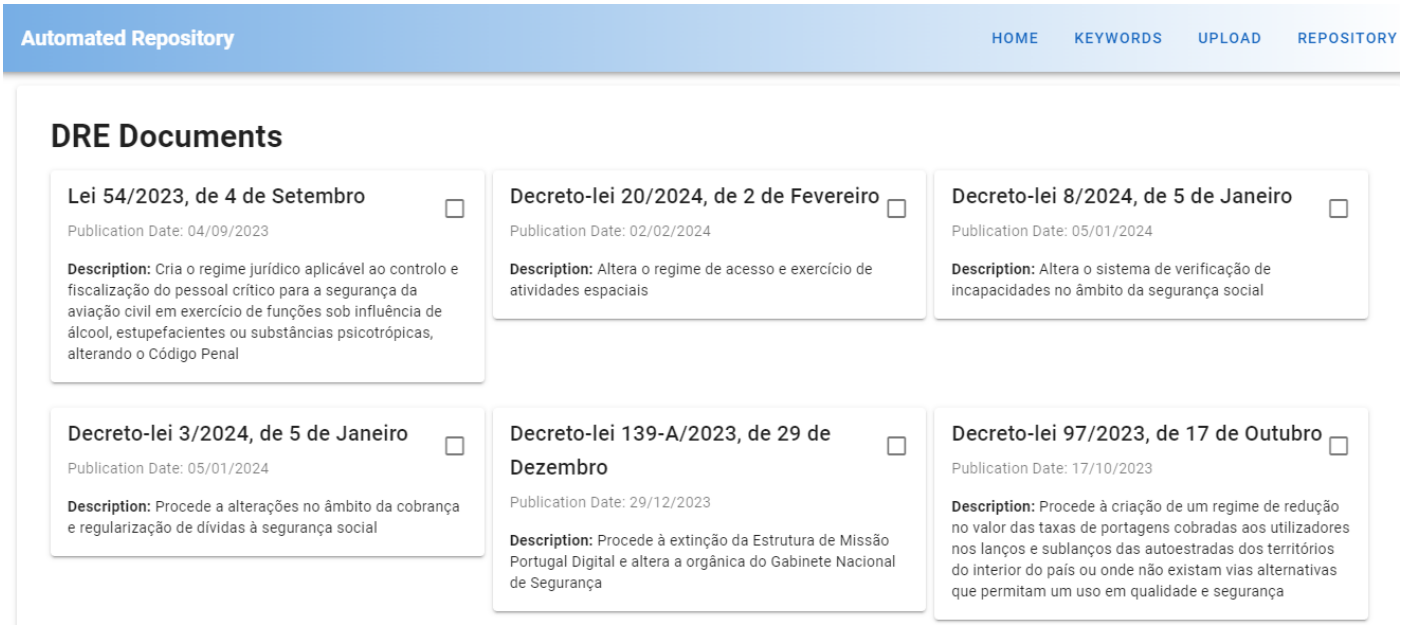

Figure 18. New Documents page.

5. Discussion

During the period between March 2024 and February 2025, a total of 582 documents were obtained from "Diário da República" [53], "Eur-Lex" [50], ENISA Publications [46], and the CNCS [35]. The documents were identified using two different methods:

- Using the tools available in the repositories to locate pages containing potential documents of interest and then submitting the page links to PDFCrawler;
- Consulting the "New Documents Page", identifying potentially interesting documents, and then sending them to PDFCrawler using the user interface button of the same page.

A total of 391 documents were classified as being relevant for our repository, while 183 were rejected by GPT and eight documents were identified as duplicated. Of the 391 relevant documents, a total of 104 documents were incorrectly identified as relevant, representing about 27% of the documents deemed relevant and 18% of all documents analyzed. In the end, a total of 287 relevant documents were made available in the repository. The percentages of incorrectly classified documents could have been lower due to two factors: the use of GPT-3.5 Turbo for a significant portion of the analyses to reduce costs, as GPT-4o mini was released after most documents had already been added to the repository, and some documents from the "Diário da República" being poorly formatted.

During the identification of relevant and non-relevant documents, it is worth noting that GPT sometimes made mistakes, as it misclassified documents with minor references to IT-related topics as being relevant to the theme, indicating a difficulty in determining the true importance of these mentions. However, it is also worth highlighting that GPT demonstrated a strong capability to produce comprehensive and accurate summaries of the documents. This proficiency often allowed the correction of errors made in other aspects of document classification.

After distinguishing between relevant and non-relevant documents, GPT classified each document and different behaviors were observed depending on the attributes generated.

- Title: As expected, most titles generated by GPT matched the official document titles. However, some titles were excessively long, and many were in all capital letters, reducing their visual appeal. Prompt refinement can address this issue.
- Subject: GPT accurately identified document subjects with minimal issues. Occasionally, it confused terms meant for the categorization of areas with those meant for subjects, leading to some inaccuracies. Again, some prompt refinements could easily prevent most of these occurrences.
- Areas: GPT performed well in straightforward cases but often labeled areas as "general" when interpretation or context was required.
- Related Documents: This attribute proved challenging, as GPT only identified documents as related if they were mentioned in the first 850 tokens provided for analysis, resulting in many documents lacking references in this category.
- Abstract: GPT excelled in generating concise yet informative abstracts that captured the documents' key insights. Interestingly, some abstracts mentioned related documents not identified in the "Related Documents" attribute.
- Other Attributes: GPT had no significant issues generating the "issuer", "origin", "type", and "date" attributes.

Although AI, and particularly GPT, is not perfect, it proved to be an invaluable tool in populating the repository. It helped not only in identifying relevant documents but also in providing varied information about each one.

Although some manual refinement was necessary for the documents identified as relevant and for certain details provided about each document, the use of GPT saved significant time and effort. Additionally, the information given by the model about each document, and particularly the abstracts generated by GPT, greatly facilitated the manual refinement process.

With these observations in mind, it can be concluded that the development of automated repositories capable of integrating AI tools, such as GPT, to classify and provide insights about documents and highlight relations between documentations can bring many advantages to areas where analyzing vast amounts of extensive documentation is necessary.

Our automated repository utilizes GPT for document analysis, complemented by human moderation and corrections, and incorporates visualization tools, schemes, and diagrams. These combined functionalities are valuable across all areas that typically handle extensive documentation, which can be cybersecurity but also areas such as law, healthcare, the natural sciences, engineering, the social sciences, and many other sectors. They are particularly beneficial for professionals conducting document searches and analyses in diverse and complex repositories. The key functionalities present in the developed repository that could be of great use for these professionals are the following.

- AI-driven document selection: GPT saved tremendous effort in detecting documents that were not of interest in the area of study. Although many repositories allow for advanced query generation, these queries require user learning and expertise. Moreover, keyword-based queries often return documents that reference the keyword but are beyond the user's context of interest. GPT selection proved to be a useful complementary filter, saving much effort in selecting documents of interest.
- Extracting basic document information: The titles, dates, and issuers of hundreds of documents were extracted with minimal effort and almost no human intervention.
- Dividing and organizing documents by area and subject: With some human refinement, the results given by AI were organized effectively.

- Producing valuable abstracts: No human interaction was needed to generate these summaries, allowing for a basic understanding of the content and objectives of the individual documents. Abstracts were generated for hundreds of documents in minutes, which is a very efficient result.
- Displaying relationships between documents: A network graph was created to show these relationships. Although some human intervention was needed, the effort was minimal, and this functionality provides valuable insights into the general landscape of the area's documentation.
- Creating visual graphs: These graphs offer a different and statistical point of view of the current state of documentation, helping professionals to understand the temporal evolution of their areas of work and activity.
- Automated document collection: Developed with the help of PDFCrawler and adapted to our project, this tool removes much of the effort needed in gathering large volumes of documents. This can save time when searching for and downloading documents. The keyword feature, adaptable to any need, also saves effort and time by pre-selecting documents based on the provided keywords.
- Dedicated page for recent documents: This page notifies users of possible documents of interest that have been recently published, allowing professionals to stay updated and informed about the latest developments and to keep their documentation up to date.

For future work, it will be necessary to adapt the repository to accommodate newer and more advanced AI models, which will be capable of providing even better and more precise outputs for the given documents.

Refining the queries made to the models for each individual document is also crucial in improving the quality of the results. This refinement can be achieved in two ways: by extending the length of the excerpts taken from each document and by enhancing the questions posed to GPT. Improvements can include asking for more details, adjusting the level of restrictions, and exploring other query optimization techniques. As GPT works based on natural language, this prompt refinement process can be achieved without significant learning effort, relying instead on the careful observation and analysis of the results. This makes the process accessible and manageable, enabling continuous improvement through iterative adjustments and evaluations.

The final aim of automated documentation repositories is to offer a more accessible and informative perspective on a particular area of study. This feature of the repository provides professionals and users with valuable insights about the documentation that was aggregated, simply by browsing the repository and analyzing the results provided by the visualization tools.

Author Contributions: Conceptualization, J.F. and M.A.; methodology, J.F. and M.A.; software, J.F.; validation, J.F.; formal analysis, J.F. and M.A.; investigation, J.F.; resources, J.F. and M.A.; data curation, J.F.; writing—original draft preparation, J.F. and M.A.; writing—review and editing, J.F. and M.A.; visualization, J.F.; supervision, M.A.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no confict of interests.

References

1. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. In *Machine Learning and the City: Applications in Architecture and Urban Design*; Wiley: Hoboken, NJ, USA, 2022; pp. 535–545.
2. Chng, S.; Lu, H.Y.; Kumar, A.; Yau, D. Hacker types, motivations and strategies: A comprehensive framework. *Comput. Hum. Behav. Rep.* **2022**, *5*, 100167. [CrossRef]
3. Wang, C.; Zhang, N.; Wang, C. Managing privacy in the digital economy. *Fundam. Res.* **2021**, *1*, 543–551. [CrossRef]
4. Quach, S.; Thaichon, P.; Martin, K.D.; Weaven, S.; Palmatier, R.W. Digital technologies: Tensions in privacy and data. *J. Acad. Mark. Sci.* **2022**, *50*, 1299–1323. [CrossRef] [PubMed]
5. Lindgren, I.; Madsen, C.Ø.; Hofmann, S.; Melin, U. Close encounters of the digital kind: A research agenda for the digitalization of public services. *Gov. Inf. Q.* **2019**, *36*, 427–436. [CrossRef]
6. Usman, M.; Felderer, M.; Unterkalmsteiner, M.; Klotins, E.; Mendez, D.; Alégroth, E. Compliance requirements in large-scale software development: An industrial case study. In Proceedings of the Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, 25–27 November 2020; Proceedings 21; Springer: Berlin/Heidelberg, Germany, 2020; pp. 385–401.
7. Kutyłowski, M.; Lauks-Dutka, A.; Yung, M. Gdpr–challenges for reconciling legal rules with technical reality. In Proceedings of the Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, 14–18 September 2020; Proceedings, Part I 25; Springer: Berlin/Heidelberg, Germany, 2020; pp. 736–755.
8. Filipovska, E.; Mladenovska, A.; Bajrami, M.; Dobreva, J.; Hillman, V.; Lameski, P.; Zdravevski, E. Benchmarking OpenAI's APIs and other Large Language Models for Repeatable and Efficient Question Answering Across Multiple Documents. In Proceedings of the 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, 8–11 September 2024; IEEE: New York, NY, USA, 2024; pp. 107–117.
9. Törnberg, P. How to use Large Language Models for Text Analysis. *arXiv* **2023**, arXiv:2307.13106.
10. GPT-3.5 Turbo. 2024. Available online: https://platform.openai.com/docs/models#gpt-3-5-turbo (accessed on 1 July 2024).
11. GPT-4o. 2024. Available online: https://platform.openai.com/docs/models#gpt-4o (accessed on 1 July 2024).
12. Ghumade, T.G.; Deshmukh, R. A document classification using NLP and recurrent neural network. *Int. J. Eng. Adv. Technol.* **2019**, *8*, 632–636. [CrossRef]
13. Cascella, M.; Semeraro, F.; Montomoli, J.; Bellini, V.; Piazza, O.; Elena, B. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J. Med. Syst.* **2024**, *48*, 22. [CrossRef] [PubMed]
14. Merchant, K.; Pande, Y. NLP Based Latent Semantic Analysis for Legal Text Summarization. In Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 19–22 September 2018; pp. 1803–1807. [CrossRef]
15. Feyisa, D.W.; Berihun, H.; Zewdu, A.; Najimoghadam, M.; Zare, M. The future of document indexing: GPT and Donut revolutionize table of content processing. *arXiv* **2024**, arXiv:2403.07553.
16. Saka, A.; Taiwo, R.; Saka, N.; Salami, B.A.; Ajayi, S.; Akande, K.; Kazemi, H. GPT models in construction industry: Opportunities, limitations, and a use case validation. *Dev. Built Environ.* **2023**, *17*, 100300. [CrossRef]
17. Savelka, J.; Agarwal, A.; Bogart, C.; Song, Y.; Sakr, M. Can Generative Pre-trained Transformers (GPT) Pass Assessments in Higher Education Programming Courses? In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1. ACM, Turku, Finland, 7–12 July 2023; ITiCSE 2023. [CrossRef]
18. Liu, S.; Healey, C.G. Abstractive Summarization of Large Document Collections Using GPT. *arXiv* **2023**, arXiv:2310.05690.
19. Aladağ, F. The Potential of GPT in Ottoman Studies: Computational Analysis of Evliya Çelebi's Travelogue with NLP and Text Mining and Digital Edition with TEI. *Culture* **2023**, *5*, 7.
20. Thippeswamy, B.; Ramachandra, H.; Rohan, S.; Salam, R.; Pai, M. TextVerse: A Streamlit Web Application for Advanced Analysis of PDF and Image Files with and without Language Models. In Proceedings of the 2024 Asia Pacific Conference on Innovation in Technology (APCIT), Mysuru, India, 26–27 July 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
21. Vallabhaneni, U.; Wutla, Y.; Dichpally, T.; Ch, V.R.R.; Gone, M.R.; Kumari, P.L. Mining Mate: A Chat Bot for Navigating Mining Regulations Using LLM Models. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 14–15 March 2024; IEEE: New York, NY, USA, 2024; Volume 1, pp. 888–892.
22. Lin, L.H.M.; Ting, F.K.; Chang, T.J.; Wu, J.W.; Tsai, R.T.H. GPT4ESG: Streamlining Environment, Society, and Governance Analysis with Custom AI Models. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), New Taipei, Taiwan, 19–21 April 2024; IEEE: New York, NY, USA, 2024; pp. 442–446.
23. Feng, Y. Semantic textual similarity analysis of clinical text in the era of llm. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, 25–27 June 2024; IEEE: New York, NY, USA, 2024; pp. 1284–1289.

24. Ibrahim, I.M.; Attallah, M.S.; Abdel Hamid, S.O.; Zween, S.T.; Abuhadrous, I. Leveraging Large Language Models for Document Analysis and Decision-Making in AI Chatbots. *Adv. Sci. Technol. J.* **2025**, *2*, 1–16. [CrossRef]
25. Litaina, T.; Soularidis, A.; Bouchouras, G.; Kotis, K.; Kavakli, E. Towards llm-based semantic analysis of historical legal documents. In Proceedings of the SemDH2024: First International Workshop of Semantic Digital Humanities, co-located with ESWC2024, Hersonissos, Greece, 26–27 May 2024.
26. Bouzid, S.; Piron, L. Leveraging Generative AI in Short Document Indexing. *Electronics* **2024**, *13*, 3563. [CrossRef]
27. Mao, Q.; Dabrowski, A.; Wei, F.; Olson, E.; Neary, R.; Yang, J.; Qin, H.; Huber-Fliflet, N. Comparative Analysis of LLM-Generated Event Timeline Summarization for Legal Investigations. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; IEEE: New York, NY, USA, 2024; pp. 4743–4750.
28. Merilehto, J. From PDFs to Structured Data: Utilizing LLM Analysis in Sports Database Management. *arXiv* **2024**, arXiv:2410.17619.
29. Wiest, I.C.; Lessmann, M.E.; Wolf, F.; Ferber, D.; Van Treeck, M.; Zhu, J.; Ebert, M.P.; Westphalen, C.B.; Wermke, M.; Kather, J.N. Anonymizing medical documents with local, privacy preserving large language models: The LLM-Anonymizer. *medRxiv* **2024**.[CrossRef]
30. Assembleia da República. 2025. Available online: https://www.parlamento.pt (accessed on 30 January 2025).
31. Governo de Portugal. 2025. Available online: https://www.portugal.gov.pt/pt/gc24/primeiro-ministro,organization=RepÃžblicaPortuguesa (accessed on 31 January 2025).
32. European Council. 2025. Available online: https://www.consilium.europa.eu/en/european-council (accessed on 1 February 2025).
33. European Commission, Official Website. 2025. Available online: https://commission.europa.eu/index_en (accessed on 1 February 2025).
34. European Parliament. 2025. Available online: https://www.europarl.europa.eu/portal/en (accessed on 1 February 2025).
35. CNCS—Centro Nacional de Cibersegurança. 2025. Available online: https://www.cncs.gov.pt (accessed on 1 February 2025).
36. ENISA. 2021. Available online: https://www.enisa.europa.eu (accessed on 1 February 2025).
37. ISO—International Organization for Standardization. 2025. Available online: https://www.iso.org/home.html (accessed on 1 February 2025).
38. *ISO-27000*; Information Technology—Security Techniques—Information Security Management Systems—Overview and Vocabulary. ISO: Geneva, Switzerland, 2018.
39. Official PCI Security Standards Council Site. 2025. Available online: https://www.pcisecuritystandards.org (accessed on 1 February 2025).
40. No Direito Português, Qual a Diferença Entre Uma lei, Um Decreto-lei e uma Portaria? 2025. Available online: https://ffms.pt/pt-pt/direitos-e-deveres/no-direito-portugues-qual-diferenca-entre-uma-lei-um-decreto-lei-e-uma-portaria (accessed on 1 February 2025).
41. Infopédia. regulamento—Infopédia. 2025. Available online: https://www.infopedia.pt/apoio/artigos/$regulamento (accessed on 1 February 2025).
42. Tipos de Legislação | União Europeia. 2025. Available online: https://european-union.europa.eu/institutions-law-budget/law/types-legislation_pt (accessed on 1 February 2025).
43. Contribuidores dos Projetos da Wikimedia. Norma técnica – Wikipédia, a Enciclopédia Livre. 2023. Available online: https://pt.wikipedia.org/w/index.php?title=Norma_t%C3%A9cnica&oldid=66145026 (accessed on 1 February 2025).
44. | StandICT.eu 2026. 2025. Available online: https://standict.eu (accessed on 1 February 2025).
45. Cyber Policy Portal. 2025. Available online: https://cyberpolicyportal.org (accessed on 1 February 2025).
46. Publications. 2025. Available online: https://www.enisa.europa.eu/publications#c3=2014&c3=2024&c3=false&c5=publicationDate&reversed=on&b_start=0 (accessed on 1 February 2025).
47. National Cyber Security Strategies—Interactive Map. 2025. Available online: https://tools.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/national-cyber-security-strategies-interactive-map (accessed on 1 February 2025).
48. Country Wiki—Octopus Cybercrime Community—www.coe.int. 2025. Available online: https://www.coe.int/en/web/octopus/country-wiki (accessed on 1 February 2025).
49. DataGuidance. 2025. https://www.dataguidance.com (accessed on 1 February 2025).
50. EU Law—EUR-Lex. 2025. Available online: https://eur-lex.europa.eu/homepage.html?locale=en (accessed on 1 February 2025).
51. CNCS—Observatório de Cibersegurança. 2025. Available online: https://www.cncs.gov.pt/pt/observatorio (accessed on 1 February 2025).
52. CNCS—Quadro Nacional. 2025. Available online: https://www.cncs.gov.pt/pt/quadro-nacional (accessed on 1 February 2025).
53. Diário da República. 2025. Available online: https://diariodarepublica.pt/dr/home (accessed on 1 February 2025).
54. MongoDB: The Developer Data Platform. 2025. Available online: https://www.mongodb.com (accessed on 1 February 2025).
55. OpenAI API. 2025. Available online: https://openai.com/api (accessed on 1 February 2025).

56. Welcome to Flask—Flask Documentation (3.0.x). 2025. Available online: https://flask.palletsprojects.com/en/stable/changes/#version-3-0-0 (accessed on 1 February 2025).
57. Vuetify—A Vue Component Framework. 2025. Available online: https://vuetifyjs.com/en/#installation (accessed on 1 February 2025).
58. Vis-Network. 2025. Available online: https://github.com/visjs/vis-network (accessed on 1 February 2025).
59. Chart.js. 2025. Available online: https://www.chartjs.org (accessed on 21 May 2024).
60. ChatGPT. 2025. Available online: https://openai.com/chatgpt (accessed on 1 February 2025).
61. OpenAI. 2025. Available online: https://openai.com (accessed on 10 February 2025).
62. OpenAI Models. 2025. Available online: https://platform.openai.com/docs/models (accessed on 1 February 2025).
63. Pricing, 2024. Available online: https://openai.com/api/pricing/ (accessed on 1 March 2025).
64. SimFin. Pdf-Crawler. 2025. Available online: https://github.com/SimFin/pdf-crawler/tree/master (accessed on 1 February 2025).
65. Openai. Tiktoken. 2025. Available online: https://github.com/openai/tiktoken (accessed on 1 February 2025).