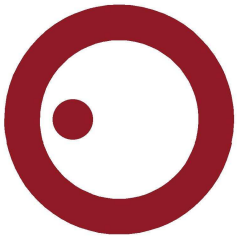


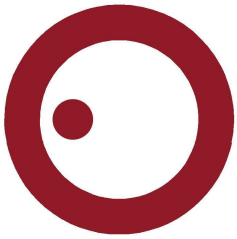
escola superior de tecnologia e gestão
instituto politécnico de leiria

Instituto Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Cibersegurança e Informática Forense

AUTOMATED CYBERSECURITY DOCUMENTATION REPOSITORY

ESTUDANTE JOSÉ MIGUEL BRAGA BARROS BAROSA FRADE

Leiria, Setembro 2024



escola superior de tecnologia e gestão
instituto politécnico de leiria

Instituto Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Cibersegurança e Informática Forense

AUTOMATED CYBERSECURITY DOCUMENTATION REPOSITORY

ESTUDANTE JOSÉ MIGUEL BRAGA BARROS BAROSA FRADE
Número: 2220723

Dissertação realizada sob orientação do Professor Doutor Mário Antunes (mario.antunes@ipleiria.pt).

Leiria, Setembro 2024

DECLATARATION

I declare, on my word of honor, that the work presented in this thesis, titled *"Automated Cybersecurity Documentation Repository"*, is original and was carried out by the student José Miguel Braga Barros Barosa Frade (2220723) under the supervision of professor Mário Antunes (mario.antunes@ipleiria.pt).

Assinado por: **José Miguel Braga Barros Barosa Frade**
Num. de Identificação: 30200616
Data: 2024.09.22 18:34:06+01'00'

Leiria, September 2024

CHAVE MÓVEL

Student José Miguel Braga Barros Barosa Frade

ACKNOWLEDGMENTS

I want to express my heartfelt gratitude to those who have supported me on this journey. To my mother, Sandra, my father, José, my sister, Marta, and my girlfriend, Bianca — your support and encouragement have been my constant source of motivation. A special thank you to my grandparents and aunts for always believing in me. I want also to thanks to my friends, who always supported me during this time. I could not have done this without you. I would also like to thank all my teachers, from kindergarten to this Master's degree, for the knowledge and guidance they have provided me over the years. A special thanks to my thesis advisor, Mário, whose knowledge and support were invaluable in completing this work. Thank you all for helping me reach this milestone.

RESUMO

A digitalização trouxe consigo inúmeras vantagens, mas à medida que a sociedade se torna mais dependente dos sistemas informáticos e também que os dados se tornam mais valiosos, os riscos no ciberespaço vão aumentando de igual forma. Para mitigar estes riscos, muitas organizações e entidades governamentais emitem documentos legais e orientações técnicas que devem ser seguidos e implementados. Embora essenciais para a cibersegurança, estes documentos podem possuir um grau de complexidade bastante elevado e a sua análise pode ser exaustiva. Além disso, o número existente deste tipo de documentação aumenta continuamente, sendo que hoje em dia podemos encontrar documentos técnicos e legais relacionados com cibersegurança dispersos por vários repositórios na Internet.

Nesta tese, desenvolvemos o "Automated Repository", que utiliza dois modelos NLP, GPT-4o e GPT-4o mini, assim como tecnologias de recolha de documentos e ferramentas de visualização para auxiliar os utilizadores a recolher, organizar, analisar e extrair informações cruciais da documentação relacionada com a cibersegurança. O desenvolvimento da aplicação é descrito ao longo deste relatório e as funcionalidades implementadas são avaliadas face aos objetivos definidos.

Com base na informação recolhida e gerada pela aplicação, foi realizada uma análise da evolução da documentação de cibersegurança ao longo do tempo, destacando-se os documentos mais relevantes e que mais marcaram a área. Por fim foi possivel concluir que o "Automated Repository" pode auxiliar significativamente os utilizadores nas tarefas de recolha e análise de documentação sobre cibersegurança. Além disso, o sistema pode ser de grande utilidade em outras áreas e setores da sociedade. Exemplos são as áreas da saúde, ciências sociais, engenharias e muitas outras onde os profissionais têm de seguir e consultar regularmente grandes volumes de documentação frequentemente.

Keywords: Automação, Cibersegurança, Documentação, GPT, Legislação, Modelos NLP, Repositório

ABSTRACT

Digitalization has introduced numerous advantages, but as society becomes more reliant on computer systems and data becomes increasingly valuable, the risks in cyberspace are increasing accordingly. To mitigate these risks, many organizations and government entities issue legal documents and technical guidelines that must be followed and implemented. Although essential for cybersecurity, these documents can be quite complex and their analysis can be exhaustive. Furthermore, the existing number of this type of documents is continuously increasing, and today it is possible to find technical and legal documentation related to cybersecurity distributed across various repositories on the Internet.

In this thesis, we have developed the "Automated Repository", which uses two NLP models, GPT-4o and GPT4o mini, as well as document collection technologies, and visualization tools to help users collect, organize, analyze, and extract crucial information from cybersecurity-related documentation. The development of the application is described throughout this report and the functionalities implemented are evaluated against the defined objectives.

Based on the information collected and generated by the application, an analysis of the evolution of cybersecurity documentation was performed, highlighting the most relevant documents and those that have influenced the area the most over time. Finally, it was possible to conclude that the "Automated Repository" can significantly assist users in the tasks of collecting and analyzing cybersecurity documentation. In addition, the system can be of great use in other areas and sectors of society. Examples are the areas of healthcare, social sciences, engineering and many others where professionals have to regularly follow and consult large volumes of documentation.

Keywords: Automation, Cybersecurity, Documentation, GPT, Legislation, NLP Models, Repository

INDEX

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence.
ANCC	National Cybersecurity Certification Authority.
API	Application Programming Interface.
BSON	Binary Javascript Object Notation.
CERT	Computer Emergency Response Teams.
CNCS	Portuguese National Cybersecurity Center.
DNS	Domain Name System.
DORA	Digital Operational Resilience Act.
DPIA	Data Protection Impact Assessments.
DRE	Republic Official Journal.
EEA	European Economic Area.
EITAS	European Travel Information and Authorization System.
ENISA	European Union Agency for Cybersecurity.
EU	European Union.
GDPR	General Data Protection Regulation.
GPT	Generative Pre-training Transformer.
ICT	Information and Communication Technologies.
ID	Identification.
IEC	International Electrotechnical Commission.

ISMS	Information Security Management System.
ISO	International Organization for Standardization.
IT	Information Technologies.
JSON	JavaScript Object Notation.
LLM	Large Language Model.
NIS1	Network and Information Security 1.
NIS2	Network and Information Security 2.
NIST	National Institute of Standards and Technology.
NLP	Natural Language Processing.
PCI DSS	Payment Card Industry Security Standards Counci.
PDF	Portable Document Format.
PSIRT	Product Security Incident Response Teams.
REST	Representational State Transfer.
RSS	Really Simple Syndication.
SQL	Structured Query Language.
UI	User Interface.
URL	Uniform Resource Locators.
XML	Extensible Markup Language.

1

INTRODUCTION

1.1 MOTIVATION

The digitalization of public and private organizations is rapidly advancing. Moreover, Information Technologies (IT) are increasingly becoming an integral part of everyone's lives. Also, services are rapidly moving to digital forms, and new technologies, such as Artificial Intelligence (AI), are rapidly evolving and gaining ground and applications in various sectors [1].

The rapid digital transformation of the world has brought numerous advantages, such as enhanced accessibility to information and services, and improved efficiency in many tasks and operations. However, it also poses some risks and challenges. As IT assets and personal data increase in value, they attract attackers from around the globe with various malicious intentions [2]. Moreover, with digitalization, concerns about abuses of personal privacy and the collection and analysis of personal data by companies and even governments have become well-documented issues [3, 4].

In addition, as we move toward an increasingly digital society, it is critical to ensure the quality and reliability of new technologies and systems used in critical sectors such as healthcare, energy, water, transportation, and more. Ensuring that this digitalization process is done in a way that does not prevent anyone from being able to take benefit from the now-digitized services is also a concern [5].

In order to combat the aforementioned problems, various governmental entities are issuing legal documents that are designed to guide a wide variety of organizations in their digital transformation process in a very diverse set of ways. Some international renowned institutions are also responsible for issuing technical standards and guidelines that, while not having legal authority, provide essential measures, policies, and guidelines. If implemented, the recommendations posing in these documents will ensure a wide set of benefits, such as improved resilience to cyberthreats, improved cyber-related policies, and management procedures. In some cases, this technical standards and guidelines implementation are also required for organizations to receive recognition and sometimes legally required certifications.

Many of these documents are accessible through on-line platforms and repositories across the Internet, but managing these governance and legal resources can be challenging due to the high volume of daily publications, the diversity of repositories, and the variety of publishers.

Navigating through different platforms, each with unique interfaces and search functionalities, can also be complex and a time-consuming task. Moreover, the rapid evolution of cyber-related norms, laws, and guidelines means that information must be continuously updated, requiring users to frequently adjust their search strategies and stay informed with new developments provided by the publishers.

Another problem is that many of these repositories contain not only documents related to cyberspace and cybersecurity, but also those related to a wide range of fields and sectors present in society.

Also, analyzing and implementing technical and legal documentation presents significant challenges, not only in cybersecurity and Information and Communication Technologies (ICT) management, but also across a wide range of sectors such as healthcare, various branches of engineering (civil, mechanical, electrical, etc.), social sciences and many more. The complexity arises from the diverse and intricate nature of the documents involved, which often represent highly specialized standards, regulations, and methodologies specific to each field. This complexity is further compounded by the need to understand and integrate these documents within their specific contexts, which can vary widely even within the same sector. As a result, accurately interpreting and applying these documents requires a deep understanding of both the technical and legal nuances unique to each domain.

Implementing legal documents and technical guidelines in organizations often requires significant effort and expertise to simplify the complexity of the document's content and understand its requirements. This analytical complexity increases when professionals need to analyze a large volume of field-specific documentation [6, 7].

In addition to all the challenges listed above, it is also important to note that the great volume and dispersion of cybersecurity legal documentation and its complexity means that many users of the digital world are unaware of their rights and legal protections regarding their personal data and rights as consumers of digital services and platforms.

The complexity of collecting, processing, labeling and analyzing documents can be significantly reduced by using emerging text ingestion tools, such as the latest Natural Language Processing (NLP) models. These models can easily abstract and explain large amounts of text in seconds, allowing for a quicker and easier

understanding of the main objectives of the document and its areas of application. This advancement in NLP technology facilitates more efficient and effective document analysis, saving time and reducing the effort required from professionals.

NLP models, like Generative Pre-training Transformer (GPT) 3.5 [8] and GPT-4o [9], use advanced techniques such as tokenization, context-aware embeddings, and attention mechanisms to process and summarize large amounts of text. The tokenization process breaks text into smaller units like words or phrases, while context-aware embeddings represent these units in a way that captures their meaning based on surrounding words. Attention mechanisms help the model to focus on the most relevant parts of the text, improving understanding and accuracy in summarization.

By identifying key themes, concepts, and relationships within the text, these models are able to generate concise summaries that capture the essence of complex documents. This innovation is very useful and important, as it allows deep learning algorithms to transmit accurate information that is easier to understand, drastically reducing the time and effort required to comprehend and apply detailed information.

1.2 GOALS

Many organizations struggle to find the specific documents necessary for compliance with laws, technical norms, and other necessary documentation and guidelines. Similarly, many individuals lack a complete understanding of their digital rights, largely because there is no single repository that consolidates all documents outlining their rights and freedoms in the digital world. Also, professionals from many sectors, including cybersecurity, often struggle to comprehend and analyze complex legal and technical documentation, a challenge that is aggravated by the increasing volume of these types of documents.

The main objective of this dissertation is to develop an application that addresses the identified needs, and to test it with cybersecurity-related documentation from both Portugal and the European Union (EU). To achieve this, the dissertation will focus on the following goals:

1. To develop a repository that aggregates legal documents and technical cybersecurity guidelines from various sources within Portugal and the EU.
2. To equip the repository with search and visualization capabilities to facilitate access to the documents.

3. To design the repository to be user-friendly and intuitive, allowing users to easily navigate and understand the documents.
4. To develop a system that heavily relies on automation to efficiently manage a large volume of documents.
5. To equip the repository with NLP models that can extract important information and filter, categorize and label each document.
6. To provide insights and conclusions of the current state of cybersecurity and cyberspace governance documentation in Portugal and EU, based on the analysis of the repository information.
7. To develop tools capable of informing users about the most recent Portuguese and European documents that might be of interest to the automated repository
8. To develop a system that could be used, not only within the context of this thesis, cybersecurity documents from Portugal and EU, but that is capable of being used in other contexts and scenarios

1.3 CONTRIBUTIONS

All code developed during this project is available on the Automated Repository's GitHub page: https://github.com/JoseMiguelFrade/Automated-Repository/tree/main. On the page, there is also a list of requirements to run the application, some recommendations, two installers (one for Windows and one for Linux) and two launchers (one for Windows and another for Linux). Lastly, a download link is available in the GitHub repository, providing a zip file that contains a complete copy of the project's database of documents and classifications. It is not necessary to download the zip file to test the application. The zip file is 616 MB, and the three uncompressed files in the database total around 1 GB. The files are in JSON format and can be uploaded to a MongoDB-managed database. Each file represents a unique MongoDB collection.

To our knowledge, no other study or application has been published that is similar to or aims at the same objectives identified in Section 1.2 as the "Automated Repository" developed.

1.4 DOCUMENT STRUCTURE

This thesis is structured as follows:

Chapter 2 provides the background to the theme of cybersecurity documentation and the usage of NLP models in the context of documentation analysis. This chapter presents concepts related to categories of legal and technical documents, enumerates the main organizations that issue cyberspace-related documentation, and identifies the primary documentation repositories. In addition, the main cybersecurity documents are listed and described. Finally, studies from the literature on the usage of NLP models in the analysis of textual information are enumerated and detailed.

Next, in Chapter 3, the development of the automated repository is detailed. The architecture is presented along with the technologies and methodologies applied. The final User Interface (UI) of the repository is described and all developed functionalities are detailed.

In Chapter 4, the key insights extracted from the automated repository are described. These insights are based on all information generated by the developed functionalities.

Finally, a general conclusion is presented in Chapter 5, providing an overview of the final results and observations. The path for future work and the applicability of this automated repository to other fields are also addressed.

2

BACKGROUND

Cyberlaw refers to the collection of laws, decree-laws, directives, and regulations specifically designed to address issues arising within cyberspace. This includes a broad spectrum of concerns related to cybersecurity, privacy, digital transactions, intellectual property, and access to digital resources, among others [10].

The main objective of these legal documents is to implement strategies and procedures to protect individuals, organizations, and governments from cybercrime, data breaches, and other malicious activities online, while also ensuring the respectful and ethical use of digital resources and data.

There are also technical norms and frameworks, such as International Organization for Standardization (ISO) 27000 family [11] and National Institute of Standards and Technology (NIST) [12] special publications, that provide guidelines and best practices for organizations to enhance their cybersecurity capabilities and protect their digital assets. These documents can be applied to various sectors, including but not limited to healthcare, defense, energy, telecommunications, and finance, acknowledging the unique vulnerabilities and requirements of each domain.

Given the global nature of the Internet, there are also international regulations and treaties aimed at harmonizing laws across borders to effectively combat cyber threats and foster a safer digital environment around the world. As digital technologies continue to advance and affect all aspects of our lives, these documents are crucial to managing cyberspace. They must be constantly updated to address new challenges and ensure that the rights and security of all internet users are protected.

In this chapter, some main legal and technical issuers of cybersecurity and cyberspace-related documentation will be enumerated, along with the main repositories for these documents. Related concepts, such as the types of legal and technical documents, will be clarified. In addition, some well-known cybersecurity documents (both legal and technical) will be explained. Finally, studies where NLP models have been applied to the analysis of documentation are enumerated. The main objective of these studies is to reduce the effort required by professionals and users to understand and implement the contents of complex documentations.

2.1 ISSUER ORGANIZATIONS

In the context of Portugal and the EU, there are several key entities that provide legal and technical documents and frameworks. This section aims to enumerate the main document issuers within the Portuguese and European context.

- **Legislation Issuers**: Legislation issuers are entities or individuals with the authority to create, modify, or repeal laws. This type of responsibilities typically lies with government bodies such as parliaments, congresses, or legislative assemblies at national or regional level. Examples of legislation issuers in Portuguese and European context are:
  - **Portuguese Assembly of the Republic** [13]: The legislative body responsible for creating and approving national laws, including those related to cybersecurity and digital operations in Portugal.
  - **Portuguese Government** [14]: Responsible for implementing and enforcing laws, including policies and regulations related to cybersecurity within Portugal. The government can also issue and implement decree-laws, as will be described in Section 2.2.1.
  - **European Council** [15]: An EU institution that defines the general political direction and priorities of the European Union, including overarching cybersecurity strategies and policies.
  - **European Commission** [16]: The branch of the EU responsible for proposing legislation, and implementing decisions, including cybersecurity regulations and directives.
  - **European Parliament** [17]: The directly elected legislative body of the European Union that works with the Council to adopt and amend proposed legislation, including laws related to cyberspace and cybersecurity.
- **Technical Norms and Frameworks Issuers**: Technical norms and frameworks issuers are entities that develop and establish standards, guidelines, reports, and best practices for specific industries or fields. These issuers can include international organizations, national standards authorities, industry associations, and professional societies. Their goal is to ensure consistency, safety, quality, and interoperability across products, services, and processes within their respective areas of expertise.
  - **Portuguese National Cybersecurity Center (CNCS)** is Portugal's national authority responsible for promoting cybersecurity and enhancing

the country's cyber resilience. Its main functions include monitoring and responding to cyber threats, raising awareness, providing expert guidance, collaborating with international partners, and supporting regulatory compliance [18].

- **European Union Agency for Cybersecurity (ENISA)** is the central cybersecurity authority for the EU, focused on enhancing cybersecurity across member states. ENISA's key activities include providing expert advice on security issues, aiding in the development of EU policy and law, promoting cybersecurity awareness and education, and facilitating collaboration among EU countries to improve resilience against cyber threats [19].
- **Internation 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 [20].
- **Payment Card Industry Security Standards Council (PCI SSC)** [21] is an international organization whose main task is to ensure the security of credit and debit card transactions and protect cardholder data. This organization is the issuer of the Payment Card Industry Data Security Standard (PCI DSS), which provides a comprehensive set of requirements for enhancing payment card security and managing cardholders personal data [22].
- **National Institute of Standards and Technology (NIST)**: NIST is a U.S. federal agency that develops standards and guidelines for technology-related fields such as cybersecurity. NIST aims to ensure consistent, reliable, and effective approaches are taken by different industries in the areas covered by its standards and recommendations [12].

- **Others**: There are numerous legal and technical documentation issuers across the globe that publish cybersecurity documentation. Almost every country has a responsible governmental body or organization (exclusively dedicated or not) tasked with providing cybersecurity legislation, recommendations and guidelines, monitoring national cyberspace, and ensuring compliance with national and applicable international cybersecurity legal and technical requirements.

2.2 TYPES OF LEGAL AND TECHNICAL DOCUMENTS

Numerous legal documents are issued to govern cyberspace, including laws, decrees, regulations, and declarations. In addition to these, numerous organizations publish a variety of technical documents aimed at helping entities improve their IT infrastructures and cybersecurity capabilities. This study will delve into an analysis of these documents, focusing on the frameworks and guidelines established by Portugal and the EU. The motivation behind this focus stems from our repository, which has been meticulously compiled from documents released by both Portugal and the EU, providing a comprehensive basis for our examination.

2.2.1 *Laws and Decree-laws*

In Portugal, laws and decree-laws have equal authority. However, the key distinction lies in their approval process: laws are approved by the Assembly of the Republic, while decree-laws are issued by the Government. The decision between applying a law or a decree-law on a specific subject in the event of a conflict depends on two factors. The one chosen should be either the more recent legislation or the one that is more specifically tailored to the issue at hand [23].

2.2.2 *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 European Union.

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. [24].

If published by the EU, a regulation becomes a binding legislative act, and it must be fully applied throughout the EU [25]. So, in other words, EU regulations introduce new laws and legal procedures that should be fully implemented by all countries in the union.

2.2.3 *Directives*

Directives are documents issued by the European Union, and their objective is to provide guidelines for member countries to follow in order to achieve specific goals and objectives. 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 [25].

2.2.4 *Technical Norms and Frameworks*

Technical norms standards and frameworks, which are not legal documents but are crucial in this area, are documents issued by accredited official entities. These documents establish rules, guidelines, or characteristics for a material, product, process or service, such as the implementation of cybersecurity procedures or systems within the IT infrastructure of target organizations. Compliance with these documents, such as an ISO standard, is not mandatory unless its implementation is required by law [26].

2.2.5 *Reports*

Some official organizations also publish informative reports on the current landscape of cybersecurity in various contexts. These reports provide valuable information on the state of cybersecurity in different sectors, demographics, countries, and more. Some reports also address the current state of technologies used in various IT environments and their behavior regarding cybersecurity, allowing professionals to choose different technologies or configurations in a more informed manner.

2.3 REPOSITORIES

All resources related to cybersecurity and cyberspace regulation are scattered across various repositories of legal and technical documents from a wide range of fields. The main repositories that can be found online and that do contain relevant documents for our context are:

- **StandICT** [27]: This repository belongs to the European Standardization Observatory and contains many IT-related standards organized by different sectors.
- **Cyber Policy Portal** [28]: It is a United Nations portal that provides a comprehensive platform for consulting a wide range of documents related to the regulation of cyberspace. It also offers insights into the institutions tasked with managing cyberspace in each country. A distinctive attribute of this repository, which makes it different from others, is its inclusion of detailed information on international agreements pertaining to cooperation in the cyberspace domain. Finally, and as a downside, it is also important to note that this portal provides some incomplete and outdated information.
- **ENISA Repositories**: ENISA offers repositories containing technical documents issued by EU member countries or by ENISA itself. Two repositories stand out particularly: ENISA Publications [29], where ENISA publishes technical documents and reports covering Europe's cybersecurity overview, regulations, emerging threats, and studies related to these topics; and ENISA's National Cybersecurity Strategies map [30], which allows users to access documents pertaining to the cybersecurity strategies of all European countries. ENISA website offers many other interactive tools that allow users to explore a variety of topics related to cybersecurity in the EU.
- **Octopus Cybercrime Community** [31]: Managed by the European Council, this portal hosts an extensive repository of information on measures, policies, and legislation adopted by numerous countries worldwide to combat cybercrime and handle digital evidence. In addition, it offers information on international agreements related to this subject.
- **OneTrust DataGuidance** [32]: 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** [33]: The EUR-Lex repository is an online resource that provides access to European Union 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 the general public, by providing advanced search tools that facilitate the retrieval of specific documents and legal information.

- **CNCS Observatory [34]**: Portugal's CNCS 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, the economic impact of cybersecurity, as well as emerging threats and cyber conflicts. CNCS also publishes technical frameworks aimed at strengthening the cybersecurity capabilities of national organizations. A prime example of this effort is the "Quadro Nacional de Referência para a Cibersegurança" [35] (National Cybersecurity Reference Framework). This framework serves as a comprehensive guide for organizations to enhance their cybersecurity measures and resilience.
- **Diário Da República (DRE) [36]**: Diário da República or "Republic Official Journal" is one of 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 and by the parliament. Diário da República is divided into two series: the first series (I Série) contains laws, decrees, and other acts of general legislative nature, while the second series (II Série) includes a wide range of other official documents, such as notices, declarations, and contracts.

2.4 WELL KNOWN CYBERSECURITY DOCUMENTS - EUROPEAN CONTEXT

This section aims to provide comprehensive insights into some main cybersecurity documents issued by the EU. The documents present in this section are: Network and Information Security 1 (NIS1), General Data Protection Regulation (GDPR), Network and Information Security 2 (NIS2) and Digital Operational Resilience Act (DORA).

2.4.1 *Directive (EU) 2016/1148 - Network and Information Security 1*

NIS1 [37] aims to ensure, through a number of defined measures, that all member states have a high level of cybersecurity, especially with regard to the networks and systems used in public administration, critical service providers and digital operators.

The directive defines critical service providers as providers that are indispensable for the normal functioning of society, such as health, energy, and transport. It also

complements possible directives that have already been defined by some states to increase their cybersecurity capabilities at the national level or in specific sectors.

It should also be noted that member states continue to be able to apply specific laws based on national contexts in these matters. For example, states may require specific security conditions from the digital service operators with which they work.

States are also encouraged to apply laws and regulations to entities that are not covered by this directive, such as private sector companies. NIS1 also refers that all measures to be taken in the field of cybersecurity must take into account a cost-benefit ratio, in order to avoid excessive financial burdens on the entities involved.

This directive contains several points:

Cooperation Group and European Agency for Network and Information Security

The first approved measure is the creation of a European cooperation group, consisting of member states and ENISA, to mediate, support, and facilitate cooperation on cybersecurity issues. ENISA will provide recommendations to the commission and member states on cybersecurity measures, operators, and service providers. In addition, ENISA will coordinate exercises to test incident response capabilities during the "CyberEurope" event.

Definition of member states' Cybersecurity National Strategies

All member states must present a national cybersecurity strategy and designate at least one entity to coordinate state activities on this topic. The strategies must address the following:

- Objectives and priorities
- Governmental framework with assigned roles and responsibilities
- Measures for incident prevention, response, and recovery, and public-private cooperation
- Integration of cybersecurity practices into education and training
- Promotion of research in cybersecurity fields
- Regular risk assessments of IT infrastructures
- Listing of all participants involved in the strategy
- Maintaining a permanent point of contact for cooperation and information exchange within the union

Member states must equip responsible entities to monitor critical service providers' capabilities to:

- Share information and documentation on IT infrastructure security
- Provide evidence of implemented measures, including audit results

For digital service operators, member states must ensure they:

- Implement security measures and incident response strategies
- Communicate results of audits, assessments, and tests
- Comply with cybersecurity laws, norms, and regulations

Computer Security Incident Response Teams - CSIRTs

The directive requires the establishment of one or more Computer Security Incident Response Teams (CSIRTs) in each member state. CSIRTs must be equipped with high technical capacity and are responsible for the prevention, response, and recovery from computer incidents that may occur within their national cyberspace. Finally, all CSIRTs from all member states should be incorporated in the European CSIRT network.

2.4.2 *Regulation (EU) 2016/679 - General Data Protection Regulation - GDPR*

GDPR [38] applies to all organizations (inside or outside the union) that collect, process, or store personal data of European citizens.

GDPR is composed of seven key points:

- **Lawfulness, fairness, and transparency**:
  - **Lawfulness**: Data processing must have a legal basis, such as consent, contract, legal obligation, vital interests, public tasks, or legitimate interests.
  - **Fairness**: Processing must not harm individuals' rights and must be easily understood.
  - **Transparency**: Data processing must be clear with accessible and regularly updated privacy notices.
- **Purpose limitation**: Data must be processed for specific, legitimate purposes communicated to the data subject. Further processing is only allowed for public interest, scientific, historical, or statistical reasons.

- **Data minimization**: Only collect necessary data to fulfill the intended and communicated purpose, maintaining it only as long as needed.
- **Accuracy**: Ensure personal data is accurate and up-to-date, correcting inaccuracies promptly.
- **Storage limitations**: Keep data only as long as necessary, with time limits for erasure or periodic reviews, except for data kept for public interest, research, or statistical purposes.
- **Integrity and confidentiality**: Secure data against unauthorized processing, loss, destruction, or damage with appropriate cybersecurity measures.
- **Accountability**: Data controllers must demonstrate compliance with data protection principles through appropriate technical and organizational measures.

GDPR also includes several other important provisions and rights that are crucial for the protection of personal data within the EU.

- **Data Subject Rights**: The GDPR grants several rights to citizens, including:
  - Access to personal data
  - Rectification of inaccuracies
  - Erasure (right to be forgotten)
  - Restriction of processing
  - Data portability
  - Objection to processing (e.g., direct marketing)
  - Protection from automated decision-making
- **Data Protection Officer (DPO)**: Required when the processor is one of the following:
  - public authority
  - an entity that performs regular large-scale data monitoring
  - an organization that operates with criminal conviction data

  The DPO ensures compliance, advises on data protection, and acts as a contact point.
- **Data Protection Impact Assessments (DPIA)**: Data processors must perform a DPIA before any high-risk data processing. The DPIA must:

  - Describe processing operations and purposes
  - Assess necessity and proportionality
  - Identify risks to data subjects
  - Implement measures to mitigate risks
- **Data Breach Notification** [39, 40, 41]: Authorities should be notified within 72 hours if a breach poses a high risk to the rights and freedoms of data subjects.
- **International Data Transfers** [42]: Data transfers to outside the EU are only allowed if:
  - Adequate data protection exists in the recipient country
  - There are safeguards related to this operations like standard contractual clauses
  - Specific derogations (e.g., consent, contract necessity)
- **One-Stop-Shop Mechanism** [43]: Organizations operating in multiple EU countries can deal with only one supervisory authority.
- **Sanctions and Fines** [44]: Non-compliance penalties up to €20 million or 4% of global turnover. Fines depend on violation severity, affected data subjects, and response.

2.4.3 *Directive (EU) 2022/2555 - NIS2*

NIS2 [45] aims to enhance the security of network and information systems across the EU by expanding the scope beyond NIS1. It includes new "critical entities" and adds sectors to the "critical service providers" category, improves several aspects of NIS1, and introduces new requirements.

- **National Cybersecurity Strategies**: This point, which was also present in NIS1 is further developed in NIS2. Some new points included in NIS2 are:
  - Cybersecurity for the national ICT supply chain
  - Measures for ICT service providers
  - Requirements and certifications for ICT products and services
  - Applying Confidentiality, Integrity, and Availability (CIA) principles to the state's open internet

  - Use of cutting-edge technologies in cybersecurity measures
  - Implementation of cybersecurity in academic and research institutions
  - Promotion of good cybersecurity practices in small and medium-sized enterprises not covered by the directive

  NIS2 requires that all member states update their national cybersecurity strategies once every five years.

- **Cybercrisis Management Framework**: All Member States must establish a cybercrisis management framework assigned to a specific entity with the necessary resources to handle large-scale cybersecurity incidents. This includes preparatory measures like national cybersecurity exercises, and detailed plans involving stakeholders (public and private) and procedures for EU assistance in case of major incidents.
- **CSIRT Requirements**: NIS2 expands on NIS1's creation of CSIRTs with additional requirements:
  - CSIRTs must be located in secure locations and have a good ticket management system.
  - They must have a secondary operating location.
  - CSIRTs must have enough staff to operate 24/7 with ongoing training.

  New CSIRT responsibilities include:
  - Monitoring threats and vulnerabilities in national ICT infrastructure and critical service providers.
  - Issuing alerts about new threats to important entities and authorities.
  - Collecting and analyzing forensic data.
  - Providing vulnerability analysis on public networks without conducting penetration tests.
  - Assisting in implementing new cybersecurity information sharing tools.
  - Cooperating with private entities to achieve directive goals.
- **European Vulnerability Database**: CSIRTs are responsible for publishing detected vulnerabilities. These vulnerabilities will be stored in an ENISA-maintained database. The database will also contain information such as affected assets or services and mitigation measures.

- **Cooperation Group**: To this group, which was created in NIS1, two new observers are added: the European Service for External Action and the European Supervision Authority. Also, new responsibilities are assigned to the group by this new directive:
  - Conducting coordinated risk assessments of critical services supply chains.
  - Providing information to CSIRTs and EU-CyCLONe.
  - Discussing significant threats like ransomware.
  - Developing a biennial plan to achieve the group's objectives.
- **EU-CyCLONe**: A new entity created by NIS2, EU-CyCLONe is responsible for supporting coordinated EU response to large-scale cybersecurity incidents. It is composed by representatives from member states' responsible authorities and supported by ENISA.
- **International Cooperation**: The EU can form cybersecurity related agreements with other countries or organizations.
- **EU Cybersecurity Report**: ENISA must publish a biannual report on the state of cybersecurity in the EU. This report will also include recommendations to improve the state of cybersecurity within the member states.
- **National Governance**: Member states must ensure that important entities and critical service providers implement NIS2 measures as well as guidelines present in other standards. The State should also ensure that the personal of these entities receive cybersecurity training.
- **Digital Entities Database**: ENISA will create a database for DNS providers, cloud providers, data center providers, managed service providers (including security), online search engines, and social networks.

2.4.4 *DORA - Digital Operational Resilience Act*

DORA [46] aims to standardize the approach to digital resilience across EU financial entities, ensuring a consistent and high level of digital security within the sector.

- **Risk Management and Governance**:
  - **Risk Management Plan**: Financial institutions must develop a plan that assigns cybersecurity responsibilities to administration. The plan

should also include adequate policies, guidelines, technologies, and procedures to protect ICT assets and operations.

  - **Identification**: All risks and strategies should be documented. ICT assets must be registered.
  - **Protection and Prevention**: Continuously monitor ICT systems for anomalies. cybersecurity tools and policies shall be used to protect assets, prevent unauthorized access, ensure data protection, and guarantee business continuity.
  - **Incident Response and Recovery**: Establish and regularly test an incident response and recovery plan to quickly address and recover from cybersecurity incidents. This includes the development of a business impact analysis.
  - **Backup and Recovery**: Implement secure backup systems to recover data and systems, and regularly test them.
  - **Learning and Evolution**: Continuously learn from incidents, improve cybersecurity measures, update policies, train employees, and implement new technologies.
  - **Communication Plans**: Develop plans to inform stakeholders, employees, customers, and the public about cybersecurity incidents and their impact, ensuring transparency and trust.

- **Risk Testing**: Risk tests must be performed by independent entities. The tests must include vulnerability scanning, open source analysis, network tests, physical security tests, penetration tests, source code analysis, etc. For advanced penetration tests, this must be done at least every three years.
- **Third Party Risk Management**: Financial entities must manage ICT risks of third-party service providers. Under DORA, financial entities are only allowed to contract third parties that comply with cybersecurity norms.
- **Information Sharing**: All incidents detected must be recorded and classified according to their severity and impact. These incidents must also be reported to relevant parties and national authorities.

In an effort to harmonize the implementation of DORA's requirements, The three European Supervisory Authorities (EBA, EIOPA and ESMA), which are responsible for supervising the implementation of this regulation, will be publishing a set of

technical standards that should be implemented in order to comply with DORA's requirements.

2.5 WELL KNOWN CYBERSECURITY DOCUMENTS - PORTUGUESE CONTEXT

This section aims to provide comprehensive information on some main cybersecurity documents issued by Portugal. The documents present in this section are: Decree-Laws 46/2018 and 65/2021.

2.5.1 *Decree-Law 46/2018 - Legal Framework for Cyberspace Security*

Following NIS1, Decree-Law 46/2018 [47] was enacted to apply the directive's aspects in Portugal. This decree-law defines rules and procedures for public administration entities, digital service operators, critical service providers, and critical infrastructure operators. Key points include:

- **National Security Strategy for Cyberspace**

  The national cyberspace security strategy consists of guidelines and measures to ensure cybersecurity in Portugal, following what was defined in NIS1 for this point. Approved by the Council of Ministers, it is evaluated by the Superior Council of National Cyberspace Security.

- **Superior Council of National Cyberspace Security**

  This council advises the government on national cyberspace security. It comprises representatives from various ministries, CNCS, and other relevant entities. In addition to helping develop and evaluate the national cybersecurity strategy, the council monitors the national cybersecurity context and develops an annual report on the strategy's execution status.

- **National Cybersecurity Center - CNCS**

  This decree designated CNCS as the national cybersecurity authority. The CNCS has the authority to regulate, supervise, monitor, and sanction within its assigned competencies. It is responsible for responding to, preventing, and recovering from incidents that affect public administration entities, digital service operators, and critical service providers. Other functions include:

  - Publishing guidelines and recommendations for good cybersecurity practices
  - Defining the national level of cybersecurity alert
  - Approving or disapproving new cybersecurity legislation
  - Collaborating with entities responsible for cyberespionage, cyberdefense, cybercrime, cyberterrorism, and data protection
  - Collaborating with any public or private entities as needed

- **Computer Emergency Response Team - CERT.PT**

  Integrated into the CNCS, CERT.PT is responsible for managing cybersecurity incidents affecting national cyberspace. It is designated as the national CSIRT and is part of the Portuguese and European CSIRT networks. Its main functions include:

  - Monitoring incidents affecting national cyberspace
  - Coordinating incident response with affected organizations' or sectors' incident response teams
  - Activating quick response mechanisms and intervening in mitigating cyberattacks
  - Coordinating and implementing risk analysis
  - Promoting cooperation between public and private entities
  - Promoting good cybersecurity practices
  - Participating in national and international information security forums and training exercises

- **Security Requirements** Target entities are required to implement measures to manage risks and threats in their systems and networks. These measures should be appropriate to specific risks and threats and aim to prevent incidents and minimize their impact. In addition to these requirements, digital service operators must ensure the security of their systems and infrastructures, have incident management and business continuity plans, document all procedures, tests, audits, and comply with international standards.

- **Incident Notification Requirements** All entities target by this decree-law must inform CNCS of cybersecurity incidents that may have a relevant impact and occur in their systems or networks. These notifications shall contain

information that allows CNCS to determine the extent and impact of the incident.

According to the different situations and notificants, CNCS will provide the necessary support and guidance. CNCS will also act as the international point of contact with foreign entities. Finally, CNCS will publish information on reported incidents when appropriate and will require critical service providers to inform entities that rely on their services of possible disruptions caused by the incident.

2.5.2 *Decree-Law 65/2021*

Decree-Law 65/2021 [48] introduces regulations for public administration, digital service operators, critical infrastructure operators, and critical service providers, building on Decree-Law 46/2018. Its goal is to enhance the security and resilience of these entities by clearly defining expected cybersecurity measures and compliance requirements.

The CNCS is responsible for applying and supervising these measures. In addition, CNCS now serves as the National Cybersecurity Certification Authority (ANCC), which allows it to issue cybersecurity certifications. The main functions of the ANCC are to develop, evaluate and implement certifications, as well as to supervise other entities that issue cybersecurity certifications.

The organizational procedures outlined in the decree-law 65/2021 are as follows:

- **Permanent Point of Contact**: Entities must designate a point of contact for ongoing communication and information exchange with CNCS.
- **Security Manager**: Organizations must appoint a security manager responsible for managing all cybersecurity measures.
- **Asset Inventory**: Organizations must maintain a detailed inventory of all essential assets for their operations.
- **Security Plan**: Organizations must define a security plan, signed by the security manager, that includes security policies, incident notification procedures, and the identification of both the cybersecurity manager and the point of contact.

- **Annual Report**: Organizations must produce an annual report that includes all cybersecurity activities, incident statistics and analysis, recommendations, and other information.

This document also defines the security measures that organizations must implement:

- **Risk Analysis**: Organizations must conduct a full risk analysis at least once a year or when there is an alert for a new vulnerability that might affect the organization. Partial risk analysis should also be conducted when there are changes on assets or after an incident occurs. Entities must document the preparation, execution, and results of the risk analysis.

  After completing the risk analysis, the organization must implement a risk management plan that:

  - Complies with sector standards approved by CNCS
  - Follows the CNCS's National Cybersecurity Reference Framework if no sector-specific norm exists
  - Ensures prevention, management, and mitigation of identified risks
  - Ensures asset resilience and recovery or redundancy for operational continuity
  - Guarantees a good response to future incidents
- **Incident Notification**: Organizations must notify CNCS of any incident that affects their operations. Notifications should be made as soon as the incident is detected, when the incident is resolved, and finally, a detailed report must be submitted.

2.6 OTHER DOCUMENTS

This section aims to provide comprehensive insights into some well-known technical norms in our analysis context: ISO's 27000 family and the Portuguese QNRCS.

2.6.1 *ISO/IEC 27000 Family*

ISO and the International Electrotechnical Commission (IEC) developed the ISO/IEC 27000 family of standards primarily to assist organizations in securely managing their

information assets, including financial information, intellectual property, employee and customer data, and information related to third parties.

These standards provide guidelines for introducing, implementing, and maintaining an Information Security Management System (ISMS) within an organization. They aim to establish a common foundation for developing organizational security practices and techniques across many types of organizations.

The ISO/IEC 27000 family of standards includes:

- ISO/IEC 27000: Overview and vocabulary
- ISO/IEC 27001: Requirements for an ISMS
- ISO/IEC 27002: Code of practice for information security controls
- ISO/IEC 27003: ISMS implementation guidance
- ISO/IEC 27004: Measurement of ISMS effectiveness
- ISO/IEC 27005: Information security risk management
- ISO/IEC 27010: Management of inter-sector and inter-organizational communications
- ISO/IEC 27014: Governance of information security

As mentioned, this family of standards includes technical norms aimed to specific sectors:

- ISO/IEC 27011: Guidelines for telecommunications organizations
- ISO/IEC TR 27015: Guidelines for financial services
- ISO 27799:2008, Health informatics — Information security management in health using ISO/IEC 27002

2.6.2 *National Cybersecurity Reference Framework*

This document [35], developed by CNCS in 2019, is aimed at all organizations operating digitally in Portugal. Its implementation is entirely voluntary, with the primary goal of providing a "cybersecurity guide" to organizations.

After an introduction and a presentation of a methodology designed to help organizations understand the risks they face, the framework is structured around 5 proposed measures:

1. **Identify** risks, resources, and critical points.

2. **Protect** people, processes, and technologies.
3. **Detect** threats in a timely manner.
4. **Respond** appropriately to isolate, mitigate, and resolve an attack.
5. **Recover** by restoring the full capacity of systems.

For each of these measures, the framework presents mechanisms and processes that enable their correct implementation.

2.7 SUMMARY

All documents described in Sections 2.4 and 2.5 are summarized in Figure 1. For each document, the figure includes detailed information, such as the document's main theme, a short abstract, and a list of its key points and objectives.

As can be observed in this chapter, the complexity of legal and technical documents in the cybersecurity domain is vast and multifaceted. These documents often address a wide range of issues, covering numerous technical and legal intricacies. Many of them contain multiple points and sub-points, whose comprehension and future implementation may require high technical expertise and, in some cases, knowledge in more than one field. Documents can also be exhaustive and may reference other documents that need to be addressed to fully implement the primary document. Given this complexity, the need for automated tools to assist in analyzing and managing these documents becomes clear.

Such tools can significantly aid professionals responsible for applying these documents, ensuring accuracy, efficiency, and compliance. Automation can streamline the interpretation and management processes, reducing the burden on specialists and enhancing the overall effectiveness of document management in the cybersecurity area.

2.8 USAGE OF NLP MODELS FOR ANALYZING DOCUMENTS AND LARGE VOLUMES OF INFORMATION

In recent years, NLP models and their potential applications have been the subject of many studies, such as their use in supporting the analysis and comprehension of complex textual information. Ghumade et al. [49] developed an NLP based document

Document	Issuer	Year of Issue	Theme	Abstract	Main Key Points
Directive (EU) 2016/1148 - Network and Information Security 1 - NIS1	European Parliament and Council	2016	Cybersecurity	This directive aims to ensure high levels of cybersecurity across member states, focusing on networks and systems used in public administration, critical service providers (health, energy, transport, etc.), and digital operators. The directive complements existing national cybersecurity measures and allows states to enforce specific laws based on national contexts. It encourages states to apply cybersecurity regulations to entities not covered by the directive, such as private companies. All cybersecurity measures must consider a cost-benefit ratio to avoid excessive financial burdens on the involved entities.	• Cooperation group and ENISA • Member states' cybersecurity national strategies • CSIRT establishment and CSIRT networks
Regulation (EU) 2016/679 - General Data Protection Regulation - GDPR	European Parliament and Council	2016	Data Privacy	GDPR objective is to enhance data protection for individuals within the EU. It mandates strict rules on data handling, giving individuals greater control over their personal information and imposing hefty fines for non-compliance. GDPR requires explicit consent for data collection, grants rights to access, rectify, and delete data, and ensures data portability and protection across all processing activities.	**7 Principal Points:** • Lawfulness, fairness, and transparency • Purpose limitation • Data minimization • Accuracy • Storage limitations • Integrity and confidentiality • Accountability **Other Points:** • Data subject rights • Data Protection Officer • Data protection impact assessments • Data breach notifications • International data transfers • Sanctions and fine • One-Stop-Shop Mechanism
Directive (EU) 2022/2555 - Network and Information Security 2- NIS2	European Parliament and Council	2022	Cybersecurity	NIS2 is a directive that aims to improve the security of network and information systems across the EU. It is an improvement on the previous NIS1 Directive. NIS2 expands the scope of entities that must comply with the Directive by including new types of entities in a new category of "critical entities". This new directive also includes some new points or additions to points already addressed in NIS1.	**New Points:** • Cybercrisis management framework • European vulnerability database • EU-CyCLONe • Cooperation outside of the EU • EU cybersecurity report • National governance • Digital entities database • New "critical entities" and other sectors added to "critical service providers" **Upgraded Points From NIS1:** • National cybersecurity strategies • New CSIRT requirements • Cooperation group
Regulation (EU) 2022/2554 of the European Parliament and of the Council - Digital Operational Resilience for the Financial Sector - DORA	European Parliament and Council	2022	Cybersecurity in the financial sector	DORA aims to standardize the approach to digital resilience across EU financial entities, ensuring a consistent and high level of digital security within the sector. This regulation, which is very complete, imposes technical procedures and security policies that must be implemented by the target entities.	• Risk management plan • Risk testing, including penetration tests • Third party risk management • Information sharing
Decree-Law 46/2018 - Legal Framework for Cyberspace Security	Portuguese Assembly of the Republic	2018	Cybersecurity	Following NIS1, Decree-Law 46/2018 was enacted, applying to Portugal the various aspects outlined in the european directive. This decree-law establishes the legal framework for the security of cyberspace,defining the rules and procedures to be followed by public administration entities, digital service operators, critical service providers and critical infrastructures operators.	• Portuguese security strategy for cyberspace • Superior Council of National Cyberspace Security • CNCS is defined as the national authority for cybersecurity • CERT.PT creation • Cybersecurity requirements for target entities • Incident notification requirements
Decree-Law 65/2021	Portuguese Assembly of the Republic	2021	Cybersecurity	This decree-law introduces a set of regulations for public administration, digital service operators, critical infrastructure operators, and critical service providers. It aims to deepen and clarify the cybersecurity measures and compliance requirements outlined in Decree-Law 46/2018 to enhance the security and resilience of their operations. CNCS is responsible for applying and supervising the implementation of these measures.	**Target entities are required to define:** • Permanent point of contact • Security manager • Security plan • Asset inventory • Annual report • Risk analysis • New incident notification requirments

Figure 1: Well Known Cybersecurity Documents

classification system that showed better performance results when compared with other NLP models available at the time.

Cascella et al.[50] discussed the advantages and challenges of the usage of the most recent NLP models, including ChatGPT, by professionals in the healthcare sector. One of the conclusions presented is that this NLP models provide great tools that are able to assist in *"manual curation, interpretation, and knowledge discovery within biomedical literature"*.

Merchant et al. [51] developed a NLP model capable of analyzing legal documents by summarizing and capturing key concepts. This tool helps lawyers in analyzing criminal and civil cases and is capable of analyzing single or groups of documents at once. The developed application received the approval of professionals in the legal field. This paper was conducted prior to the launch of ChatGPT.

Feyisa et al. [52] present a solution to automatically classify and index specifications related to civil construction using GPT-3.5 Turbo for textual analysis and the Donut NLP model to classify document pages and extract information from images. The final objective is to develop a JSON-structured table of contents for all documents analyzed, allowing users to more easily navigate and consult this complex documentation.

In addition, in the construction field, Saka et al. [53] conducted a deep analysis of the potential applications of OpenAI models in the sector. They identified documentation management as one of the most suitable tasks for GPT to assist professionals.

A comprehensive analysis of GPT-4's performance in helping professionals analyze complex textual information on specific topics was conducted by Savelka et al. [54]. The study focused on annotated sentences from court cases, comparing the annotations made by GPT-4 to those made by trained law students. Annotating legal cases requires significant field expertise and knowledge. The authors found that GPT-4 performed acceptably in these specialized tasks but noted that there is still room for improvement. They also discovered that using batch predictions, instead of analyzing documents one at a time, is more cost-effective, though it comes with a small tradeoff in performance. Finally, refining prompts proved to be an effective way to enhance response quality, but on the other hand, minor changes to the prompt could also sometimes provide weaker results.

Liu et al. [55] present a solution to the token quantity limitations of most NLP models, including OpenAI's models. According to the authors, these limitations prevent the models from summarizing large volumes of extensive documents. To

address this, they use a combination of different technologies to achieve semantic clustering and reduce the total document size. By applying these techniques, large quantities of documents can be processed to fit within GPT's requirements, allowing for effective summarization by the model.

An unique application of NLP model capabilities is described by Aladağ et al. [56]. The author utilized OpenAI's GPT and other text analysis tools to examine a 17th-century Ottoman text written in an old language. Although several additional steps were identified as necessary, the author successfully used these tools to extract valuable insights from the analyzed text. This demonstrates that GPT and other NLP technologies can be valuable for historical research.

This thesis differentiates itself from all the mentioned studies by focusing on the usage of NLP, specifically OpenAI's GPT models, integrated into a system that allows individual users to collect, classify and analyze cybersecurity-related documents more efficiently. The system leverages NLP capabilities, visualization tools, and document gathering technologies to provide intuitive graphs and user interfaces to display information based on all relevant and cybersecurity related documents available in the main repositories. The focus is not only on evaluating whether NLP models can help professionals in complex document analysis, but also on developing a system that enables users to quickly and effortlessly obtain rich insights in the field of cybersecurity documentation.

3

DEVELOPMENT OF AN AUTOMATED REPOSITORY

In this chapter, the development of the Automated Repository will be presented in detail and discussed. All technologies used will be described, as well as the mechanisms implemented to create an application that addresses the objectives listed in Section 1.2.

3.1 DEVELOPMENT

The aim of this research, is to create an automation-based repository. The application shall be web-based, simple, and intuitive to use. One goal of the repository is to provide a tool that enables users to maintain an up-to-date and well-informed perspective on cybersecurity regulations, legislation, technical documents and guidelines and many more types of documentation related to this thematic.

To develop this repository, it was essential to establish an implementation structure that comprises various components, as will be discussed in the following section. Additionally, the deployment of diverse technologies and tools, each with unique capabilities, was necessary to create a repository equipped with multiple functionalities and innovative features.

This chapter will detail all the technologies and methodologies chosen to develop the repository. It also describes all the functionalities developed and explores the capabilities of the Automated Repository.

3.1.1 *Overall Architecture*

The architecture of the developed repository is depicted in Figure 2, and it consists of three main components: the frontend, which contains all visualization tools and menus, the backend that is responsible for data collection and processing, and the database that will store all repository information.

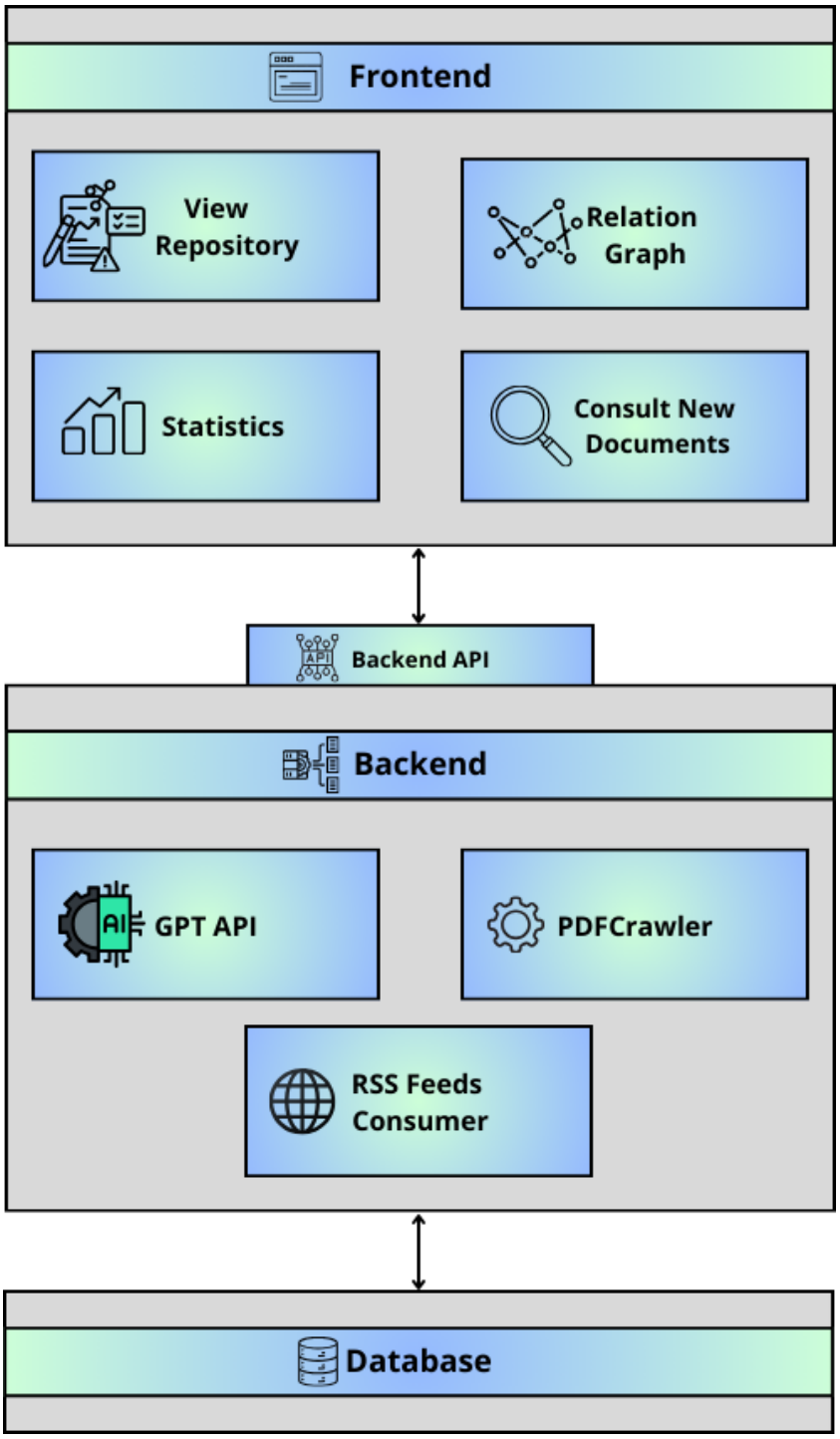


Figure 2: Repository Architecture

All documents available in the repository are stored in a database, for which MongoDB [57] has been chosen as the database management system. The rationale behind this choice will be elaborated upon in Section 3.1.3. The backend interacts with the database to store new documents or perform operations on existing ones.

The developed backend has three main software components:

- **ChatGPT API:** The utilization of the GPT Application Programming Interface (API) [58] (whose functioning details are described in Section 3.1.2) allows the utilization of a Large Language Model (LLM) type AI model for automating various tasks related to the documents. This will be further described in Section 3.2.
- **PDFCrawler:** This crawler is specialized in obtaining documents in Portable Document Format (PDF) from specified Uniform Resource Locators (URL) and is the focus of Section 3.1.4.
- **Really Simple Syndication (RSS) Feeds Consumer:** This backend module is responsible for querying external repositories in search of news documents. The implementation of this functionality is described in Section 3.1.5.

The backend is responsible for receiving instructions from the frontend, processing requests, and returning results via its API, as shown in Figure 2. The API follows the Representational State Transfer (REST) standard and was developed using the Flask framework [59].

The frontend's main task is to present the repository's information to the user in the most intuitive and comprehensive manner possible. It allows the user to manage existing documents, add new documents, and provides important insights via its "Relation Graph", which will be described in Section 3.3.2, and the "Statistics" component, which will be covered in Section 3.3.3.

The frontend was fully developed using the *"Vuetify 3"* framework [60]. The Relation Graph was implemented with the help of the *"vis-network"* library [61], and the graphs in the Statistics component were developed using the chart.js library [62].

3.1.2 *GPT API*

ChatGPT [63] is a LLM developed by OpenAI [64], designed to understand and generate human-like text based on the input it receives. It is part of the GPT series, which represents a significant leap forward in NLP technologies. ChatGPT's architecture is built upon the Transformer model, a type of deep learning model that enables a more effective handling of sequential data, such as text.

The core mechanism that allows ChatGPT to generate coherent and contextually relevant responses is its training process, which involves two main stages: 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. This learning phase enables the model to understand language patterns, grammar, and context.

Fine-tuning, on the other hand, involves training the model on smaller, task-specific datasets to adapt its responses to specific use cases. This stage is crucial for enhancing the model's performance on particular tasks, such as conversational responses, technical explanations, or creative writing. Today there are many versions of ChatGPT models available [65], being the main ones the GPT-3.5, GPT-4, corresponding "Turbo" versions, GPT-4o and GPT-4o Mini.

GPT-3.5 is a powerful language model known for its ability to generate human-like text based on the input it receives. It excels in various tasks, including conversation, text completion, and content creation. However, it may sometimes struggle with highly nuanced or specialized queries due to its limitations in understanding and context size.

GPT-4 represents a major upgrade in terms of complexity and understanding, featuring a larger number of parameters and an April 2023 knowledge cutoff. This increase allows for a deeper comprehension of context, more nuanced and accurate text generation, and improved performance on complex tasks. GPT-4 is more adept at handling intricate queries and generating sophisticated responses.

The "Turbo" version of both GPT-3.5 and GPT-4 are designed for faster response times and improved efficiency in generating text, while maintaining the remaining capabilities of both base models.

GPT-4o, with the "o" standing for "omni" has many capabilities across a wide range of tasks and contexts. It can combine text and image processing, interacts in a more human-like manner with improved natural language processing, and possesses a broader knowledge base across many areas with an October 2023 cutoff.

This version is designed to be more versatile and universally effective, handling diverse queries with improved efficiency and accuracy. Its improvements in language understanding and contextual comprehension make GPT-4o a more robust and adaptable model, suitable for a variety of applications while being cost-effective. Currently, GPT-4o is the most powerful GPT model available.

GPT-4o mini is a low-cost version of GPT-4o. Despite its lower price, it shares the same training cut-off as GPT-4o, which is October 2023, and offers better response quality comparable to its main competitor, GPT-3.5 Turbo. Furthermore, GPT-4o mini is the most recent model, having been released in May 19th, 2024.

OpenAI offers access to GPT models primarily through its web portal, as well as by providing an API that enables programmers to make requests to GPT models directly within their code.

In the case of our Automated Repository, the use of the GPT API was a key factor that enabled the utilization of GPT models' capabilities in this project.

The GPT API was used to filter and label documents, extract relevant data, and generate information based on the same documents. Initially, the GPT models used were GPT-3.5 Turbo and GPT-4.0 Turbo. After GPT-4o and its mini version were released on May 13 and 19, 2024, our Automated Repository was updated to use these models instead of GPT-4.0 Turbo and GPT-3.5 Turbo due to their advantages in terms of quality and cost. Section 3.2 further explains the implementation and usage of the GPT API in our application.

In Figure 3 an example of an API call to GPT-4o can be observed.

```
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 3: Example of a call to GPT-4o API

GPT API calls allow for the specification of several parameters, among which the most important include:

- **Model**: Within the API call, the "Model" parameter specifies which GPT model will receive and process the request. There are various models available,

such as “gpt-4o”, “gpt-4-0125-preview” (GPT-4 Turbo) and “gpt-3.5-turbo-1106” (GPT-3.5 Turbo). The versions of the models and their respective designations are updated regularly. Information about model versions and their names can be referenced in [66]. This parameter is mandatory.

- **Messages**: This parameter contains the prompts intended to be processed by the model. A message comprises two parts: “role”, which is used to provide context to the model, and "content", which contains the message itself to be sent to the model. Although not mandatory, it is recommended that the first message be sent with the role set to “system” ("role": "system") and should specify in the content that the system/GPT model should behave like an assistant ("content": "You are a helpful assistant"). After this initial message, subsequent messages can assume the context of a user making requests to the system (a helpful GPT model assistant). This parameter is mandatory.
- **Temperature**: The “temperature” parameter is a value ranging from 0 to 2 that influences the model’s behavior in responding to a request. Lower temperature values prompt the model to respond in a more factual and precise manner, whereas higher values signal the model to be more creative with its responses. This parameter is optional.
- **Top_p**: “Top_p” is a parameter that, although it enables users to adjust various specific values within the model’s functioning, ultimately shares the same purpose as the “Temperature” parameter: it allows control over the model’s technical or creative behavior. It is recommended that, ultimately, only one of the parameters (either “Temperature” or “Top_p”) be adjusted away from its default values. The default value of “Top_p” is 1.0. This parameter is optional.
- **Max_tokens**: Tokens in the GPT context are the basic units of text (such as words, parts of words, or punctuation) that the model processes and generates during training and interaction. A token can correspond to approximately 3 to 6 letters, depending on the word, its punctuation, and other factors. The parameter “Max_tokens” specifies the maximum number of tokens the model can output in its response. This value is particularly important for considerations related to pricing, as will be described subsequently. This parameter is optional, but it is highly recommended to be defined.

Finally, when discussing the GPT API, it is crucial to address pricing. Although some GPT features are available for free through the web portal, such as using the

GPT-3.5 Turbo and GPT 4o, all models incur charges when accessed through the API.

Pricing is calculated based on the tokens inputted, i.e., the number of tokens used to construct each message sent to GPT, plus the number of tokens outputted by the model in response to the request. Although we can directly control the number of tokens sent through the API by writing shorter or longer requests, the only way to manage the number of output tokens is to set the parameter "Max_tokens" to establish a maximum limit.

The final price will depend on the number of input tokens, the number of output tokens, and the model used. The price per token is always cheaper for input tokens. Regarding the models pricing, it is incorrect to assume that stronger models will be more expensive to use. In fact, GPT-4o API is cheaper than the GPT-4 and GPT-4 Turbo versions. Also, GPT-4o mini is currently the cheapest model available, offering better cost and text generation capabilities than GPT-3.5 Turbo [67].

3.1.3 *MongoDB*

MongoDB [57] is an open source, non-Structured Query Language (SQL) database managment system that uses a document-oriented approach to store data in flexible Javascript Object Notation (JSON) format documents. It is designed for scalability and handling diverse data types. For storing large files, MongoDB utilizes GridFS, a specification for storing and retrieving files that exceed the Binary Javascript Object Notation (BSON) document size limit of 16MB. 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 our project, as it is the technology currently used to store all the documents available in the repository.

3.1.4 *PDFCrawler*

One of the primary objectives in building this repository was to emphasize the provision of automation tools to users, aiming to simplify tasks such as the collection of new documents for inclusion in the repository. Among these tools is PDFCrawler.

The PDFCrawler, initially developed by [68] and later adapted for our project, enables users to input URLs to repositories, such as those mentioned in Section 2.3, and retrieve all PDFs that contain specific keywords from a user-configurable list.

To ensure a complete analysis of the repositories, the PDFCrawler employs a headless web browser (GECKO) to render the websites and then automatically simulates clicks on all clickable elements present on the visible page. Additionally, PDFCrawler allows users to set the search depth. The minimum depth value is 1, indicating that only the current page to which the given URL leads will be crawled.

For higher depth values, the crawler can behave in two separate ways, depending on the user choice. Users can choose to operate the crawler in two separate ways: "Crawl in Depth" or "Crawl in all Directions".

The "Crawl in Depth" option restricts the crawler to only crawl sub-pages of the initially provided page. For example, if the initial URL is */example*, the crawler will only crawl URLs that follow the pattern */example/...*.

The "Crawl in All Directions" option 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 option provides a more comprehensive analysis of the website, but also increases the complexity of the task. As the depth values increase (3 or more), the time and resources required for crawling can increase significantly.

Finally, PDFCrawler allows users to define a keyword list. The keyword list should contain words that the crawled PDFs must include to be stored. For each PDF crawled, PDFCrawler compares its content against the provided keyword list. If none of the keywords are present in the PDF, the file is discarded by the crawler, and the crawling operation continues. This keyword list acts as a "prefilter" for our Automated Repository, preventing all PDFs from a repository, potentially containing dozens or hundreds of files, from being stored unnecessarily. This approach improves storage management and reduces the final number of queries made to the GPT API.

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 it processes.

3.1.5 *RSS Feeds Consumer*

RSS is a web feed format used to publish frequently updated content such as blog posts, news articles, and other online content in a standardized format, typically

XML. The option to query documents using an RSS feed was available in two of the repositories used in this project: DRE and EurLex, each with its unique characteristics and specifications.

RSS Feeds Consumer ultimately allows the user to stay updated and informed about the most recent cybersecurity documents that may be of interest to both the user and the repository. If the user selects documents from the RSS Feeds Consumer, the documents links are sent to PDFCrawler. The crawler will then retrieve the documents, allowing the continuation of the addition of these PDF files to the Automated Repository.

3.1.5.1 *DRE RSS Feed*

Although DRE provides its own RSS feed, it is only available via email and lacks customization options. While this is not a significant impediment, a more user-friendly and customizable approach is available.

The website *dre.tretas.org* [69] provides an RSS feed that offers XML-structured information about all documents currently present in DRE. It also allows customization of the information in the RSS feed through URL queries, which greatly simplifies the refinement work needed to extract the necessary data from the feed.

The queries developed to gather information about recent DRE documents using *dre.tretas.org* are represented in Figure 4. The figure shows that the start date ("começo") and end date ("fim") are set to a period that occurred more than a month ago.

This is because the actual dates used in the queries are generated by our repository based on the current system date. The predefined dates shown in the query are placeholders, which are replaced by these generated dates. Only after this replacement is the query sent to the RSS feed, and the data is collected. In the end, the start date will be the date one month (30 days) ago and the end date will be the current day.

The other query parameters remain unchanged during this process. The parameter "tipo" (or "type" in English) indicates the type of documents we aim to obtain with the query. Since only one document type can be defined per query, and both law ("LEI") and decree-law ("DECRETO LEI") are of interest to our repository, two separate queries are executed. The results from these queries are then aggregated and displayed to the user.

Additionally, each query includes two keywords: “ciberespaço” (cyberspace) and “cibersegurança” (cybersecurity). Only documents that mention one of these keywords are retrieved, making these the final parameters in the queries used to fetch documents from DRE.

```
https://dre.tretas.org/dre/rss/?q=começo:2020-01-01 fim:2020-12-05 tipo:"LEI"
↪ "ciberespaço" "cibersegurança"

https://dre.tretas.org/dre/rss/?q=começo:2020-01-01 fim:2020-12-05 tipo:"DECRETO
↪ LEI" "ciberespaço" "cibersegurança"
```

Figure 4: RSS Query used on dre.tretas.org

An example of a response obtained from *dre.tretas.org* using a query identical to the ones presented in Figure 4 is represented in Figure 5.

```
<?xml version="1.0" encoding="utf-8"?>
<rss version="2.0" xmlns:atom="http://www.w3.org/2005/Atom">
  <channel>
    <title>Diários da República</title>
    <link>https://dre.tretas.org</link>
    <description>Modificações e novos documentos acrescentados ao
    ↪ site</description>
    <atom:link href="http://dre.tretas.org/dre/rss/" rel="self"></atom:link>
    <language>pt-PT</language>
    <lastBuildDate>Wed, 20 Dec 2023 00:00:00 +0000</lastBuildDate>
    <ttl>43200</ttl>
    <item>
      <title>Decreto-lei 116/2023, de 20 de Dezembro</title>
      <link>http://dre.tretas.org/dre/5588639/decreto-
        lei-116-2023-de-20-de-dezembro</link>
      <description>Transfere para o Centro Nacional de Cibersegurança as
      ↪ competências de fiscalização e de instrução de contraordenações no
      ↪ âmbito do ECOMPENSA</description>
      <dc:creator xmlns:dc="http://purl.org/dc/elements/1.1/">Presidência do
      ↪ Conselho de Ministros</dc:creator>
      <pubDate>Wed, 20 Dec 2023 00:00:00 +0000</pubDate>
      <guid>5588639</guid>
    </item>
  </channel>
</rss>
```

Figure 5: RSS Feed obtained from dre.tretas.org

The RSS Feed XML response will later be used to populate the Automated Repository's "New Documents Page" with information about the most recent documents that might be of interest to our automated repository. This functionality will be further explained in Section 3.3.4.

3.1.5.2 *EurLex RSS Feed*

EurLex allows users to register an account on the repository and save their queries as RSS feeds. When this option is chosen, the output of the query is made available via a URL and formatted in XML according to the RSS standard. The query results are always updated, but the query itself is not editable. If further refinement is needed, a new query must be made on EurLex, followed by the creation of a new RSS feed.

To address this refinement issue, a base query was created to return all results from the year 2024 up to the start of 2025. The results must only include documents representing Directives or Regulations and must contain the keywords "cybersecurity" or "information and communication technologies". After creating this query, it was saved as an RSS Feed.

An example of an RSS response obtained from this query is shown in Figure 6.

After the response is received by our automated repository, an extraction process is initiated prior to displaying the documents. During this process, only documents with dates within the interval between the last month and the current date are sent from the backend to the frontend.

Document information is also structured by the backend to match the format of the *dre.tretas.org* RSS feed. The information is organized into fields such as title, description, publication date, and URL. This ensures a consistent and uniform display of information by the frontend on the "New Documents Page", which will be described in Section 3.3.4.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" media="screen" href="./../screen-rss.xsl"?>
<rss version="2.0">
   <channel xmlns:dc="http://purl.org/dc/elements/1.1/">
      <title>EUR-Lex | CybersecuritySearchv2 | EN</title>
      <pubDate>Thu, 29 Aug 2024 13:20:32 +0200</pubDate>
      <description>
         Domain: All, Type of act: Regulation, Directive, Date: Date of
         ↪ document,
         From: 01/01/2024, To: 01/01/2025, Exclude corrigenda: True, Results
         ↪ containing:
         cybersecurity In title and text, OR: information and communication
         ↪ technologies
         In title and text, Search language: English, Exclude consolidated
         ↪ versions: True
      </description>
      <language>en</language>
      <link>https://eur-lex.europa.eu/homepage.html</link>
      <image>
         <title>EUR-Lex | CybersecuritySearchv2 | EN</title>
         <url>https://eur-lex.europa.eu/images/eurlex_logo.png</url>
         <link>https://eur-lex.europa.eu/homepage.html</link>
      </image>
      <item>
         <title>CELEX:32024R1772: Commission Delegated Regulation (EU) 2024/1772
         ↪ of 13 March 2024 supplementing Regulation (EU) 2022/2554 of the
         ↪ European Parliament and of the Council with regard to regulatory
         ↪ technical standards specifying the criteria for the classification
         ↪ of ICT-related incidents and cyber threats, setting out materiality
         ↪ thresholds and specifying the details of reports of major
         ↪ incidents</title>
         <description/>
         <link>https://eur-lex.europa.eu/legal-content/AUTO/?
         uri=CELEX:32024R1772</link>
         <guid>https://eur-lex.europa.eu/legal-content/AUTO/?
         uri=CELLAR:fa35ec94-328d-11ef-a61b-01aa75ed71a1</guid>
         <category>3</category>
         <pubDate>Tue, 25 Jun 2024 00:00:00 +0200</pubDate>
         <dc:creator>Directorate-General for Financial Stability, Financial
         ↪ Services and Capital Markets Union, European
         ↪ Commission</dc:creator>
      </item>
   </channel>
</rss>
```

Figure 6: RSS Feed obtained from EurLex Query

3.2 AUTOMATED COLLECTION AND CLASSIFICATION

To simplify document collection from multiple repositories, users can use PDFCrawler, as described in Section 3.1.4. Additionally, the RSS Feed Consumer, accessible via the "New Documents Page" (Section 3.3.4), keeps users updated on new documents relevant to the Automated Repository.

After collection, the gathered PDFs will be analyzed by GPT through the GPT API.

In this process, for each PDF, the initial step involves extracting the first 850 text tokens from the file. The reason for sending only excerpts of the PDF documents, rather than the entire document, primarily revolves around better managing monetary costs and minimizing potential errors with larger documents.

This is because there are limits to the number of tokens that the models can analyze (128K tokens for GPT-4 Turbo, GPT-4o and GPT-4o mini and 16k for the formerly used in our application GPT-3.5 Turbo).

Additionally, after conducting several tests, we concluded that 850 tokens are sufficient to provide the models with enough context and information about the PDF document, ensuring the quality of the responses generated.

To accomplish this, we utilized OpenAI's Tiktoken library [70], which enables the measurement of the quantity of tokens in a text sample. This is done using the specific way GPT models count and extract tokens from text. It is important to note that different NLP models may count tokens in various ways, hence the necessity to ensure that we are measuring tokens in a manner consistent with the model in use. Next, the 850 tokens extracted will be sent to a GPT model for analysis.

To optimize monetary operational costs, 6 out of 10 PDF excerpts will be analyzed using GPT-4o, while the remaining 4 will be sent to GPT-4o mini. Users can adjust both the number of PDF files analyzed by GPT at a time and the number of queries performed by GPT-4o and GPT-4o mini. Figure 7 shows the repository update page, where these settings can be modified. In the "Update Page" menu, *Total Queries* indicates the number of PDF files to be analyzed from the selected subdirectory. These subdirectories, that contain the crawled documents, are created by PDFCrawler for each unique website it crawls. *Low-Cost Queries* specifies the number of queries that will be performed using the low-cost GPT-4o mini model. The remaining queries will be processed using the GPT-4o model.

For both models the request's messages will be the one present in Figure 8.

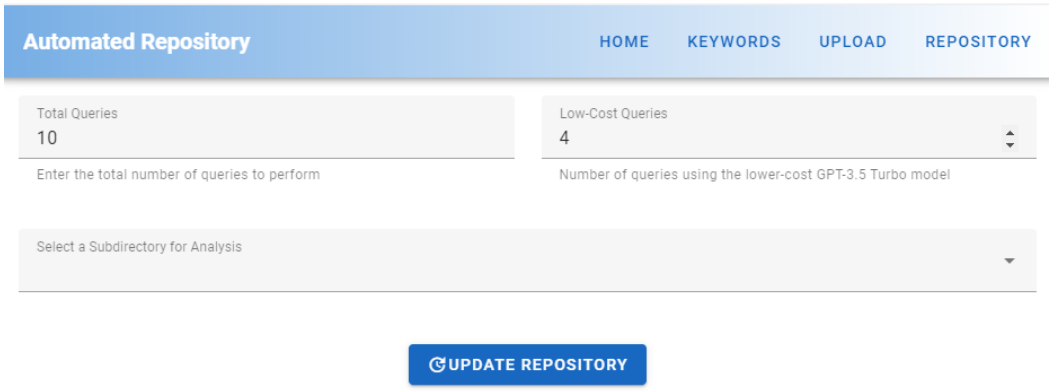


Figure 7: Update Page

```
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/Cyberecurity/...>#date:<date in dd/mm/yyyy format>
#area:<Finance/Healthcare/General/Energy/...>#title:<document_title>
#Related_Docs:doc1|doc2|doc3#abstract:<brief_summary (95 tokens max)>'."}
],
```

Figure 8: Messages sent to GPT Models

In terms of composing the messages, for each PDF document, two prompts will be sent to GPT. The first prompt will instruct the system to act as an assistant, informing it that all responses should be formatted in the following manner:

`"field:<field_value>"`

It's worth noting that the GPT API provides a parameter that allows users to inform the model, outside the prompt, that it should respond in a JSON-like format:

`"response_format={"type": "json_object"}"`

Although setting the response format as a JSON object can be useful in many situations, we found that it often caused issues in our case. These problems were

mostly related to the model's responses being restricted by their length. When the length exceeded the value of max_tokens, the final prompt would become deformed or cut off. Consequently, the final product would not qualify as a valid JSON object. Furthermore, even when successful, the final answer would consume more text tokens than the same answer using our custom format.

Following this, the subsequent message is constructed within the user's context, prompting the model to classify the PDF based on the provided excerpt: "Analyze the following document extract: 'pdf_text'".

The prompt then asks whether the document from which the excerpt was taken pertains to a topic related to cybersecurity, IT, data privacy or AI. If the answer to this question is 'no', then the model is instructed to limit its response to this question only, stopping to address subsequent inquiries. This strategy is employed to conserve output tokens. On the other away, if the answer is yes, then the model will be tasked with providing a variety of different pieces of information about the document from which the excerpt is taken:

- **Title**: The official document designation. For example, Regulation (EU) 2016/679.
- **Date**: It will be considered the date of the first publication.
- **Origin**: Where the document originated from. It can be the country, the institution, etc.
- **Issuer**: The official entity that published the document. European Parlimment, Portugal's "Presidência do Conselho de Ministros" are notable examples.
- **Subject**: The document's subject may contain one or more of the following options: Cybersecurity, intended for documents strictly focusing on cybersecurity procedures. Data Privacy is suited for documents that focus on data management and privacy concerns. Finally, Governance is intended for documents that delve more into policies, good practices, and related topics.
- **Area**: The document's area is a field designed to indicate the sector targeted by the document. Identified areas include Healthcare, Justice, Defense, Digitalization, Digital Rights, Cybersecurity, among others. The complete list of identified areas will be discussed in Section 3.3.2.
- **Related Documents**: This field may include documents identified as directly related to the analyzed excerpt. It can be empty if the analyzed excerpt does not reference any other document.

- **Abstract**: A brief summary, limited to around 95 tokens, will be generated from the provided document's excerpt. The goal of this field is to offer users of the Automated Repository a concise and straightforward description of the document.

Once the model has retrieved all the requested information, the data is stored in the database along with the corresponding PDF file, the file hash, and the date the file was uploaded to the repository.

Figure 9 represents a simplification of the process of classifying and storing documents.

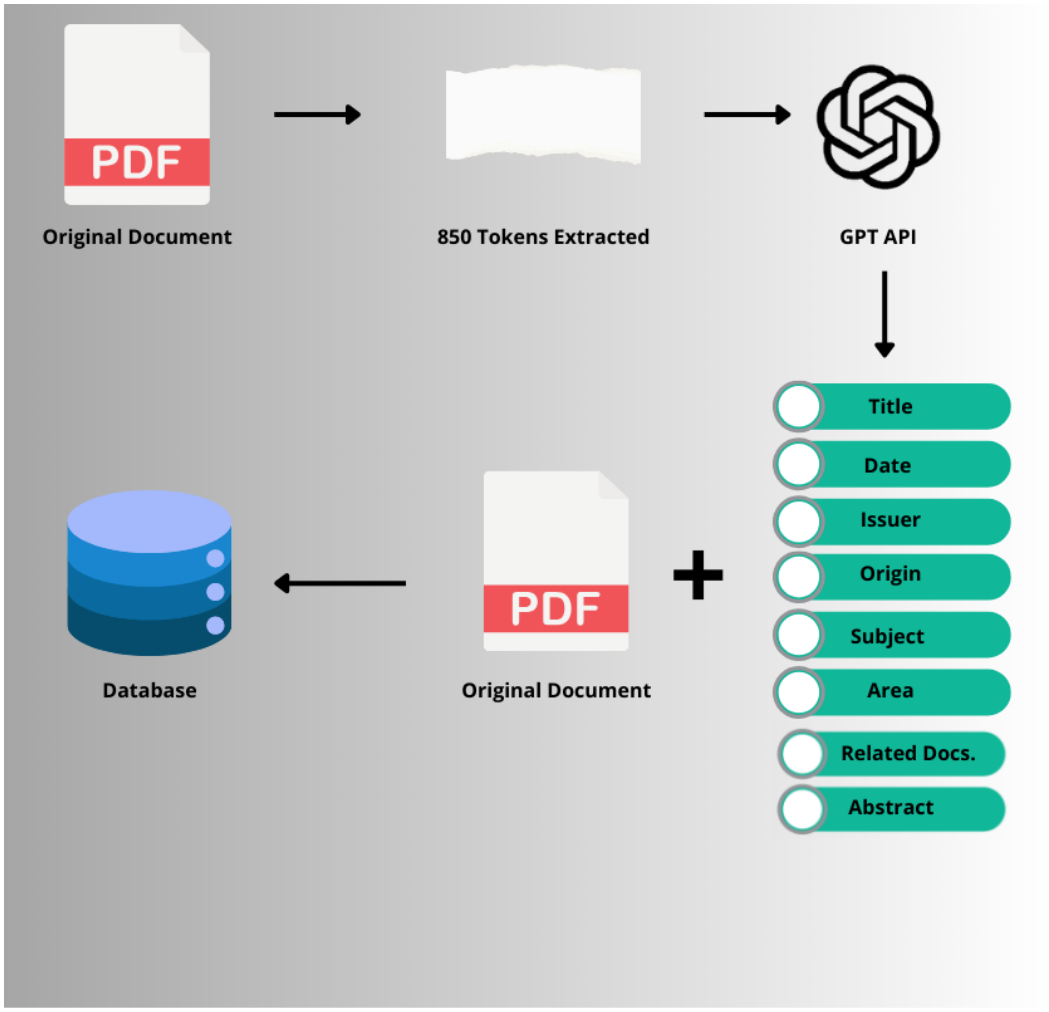


Figure 9: Simplification of Adding Documents to Automated Repository Process

As mentioned in Section 3.1.1, the database management system chosen for this project is MongoDB. The details about the document are stored in a single collection, while the PDF file itself is stored using the fs.chunks and fs.files collections.

An example of how this information is stored can be observed in Figure 10. The "pdf_file_id" value is used to fetch the corresponding PDF file from the fs collections.


```
{
  "_id": {
    "$oid": "65e6fa01803d02a35e8e21e8"
  },
  "is_related": "yes",
  "issuer": "European Parliament and the Council of the European Union",
  "origin": "European Union",
  "type": "Regulation",
  "subject": "Governance",
  "date": "27/10/2022",
  "area": "Digitalization",
  "title": "REGULATION (EU) 2022/2065",
  "related_docs": [
    "Directive 2000/31/EC"
  ],
  "abstract": "Regulation (EU) 2022/2065, known as the Digital Services Act, was
  ↪ adopted by the European Parliament and the Council on 19 October 2022. It
  ↪ aims to harmonize the conditions for providing intermediary services
  ↪ across the EU's single market by updating the legal framework established
  ↪ in Directive 2000/31/EC. This regulation addresses the challenges posed by
  ↪ new digital business models and societal risks, ensuring responsible
  ↪ online behavior and safeguarding fundamental rights like freedom of
  ↪ expression and consumer protection. Diverging national laws negatively
  ↪ affecting the internal market are also tackled through this regulation.",
  "pdf_file_id": "65e6fa01803d02a35e8e21e0",
  "pdf_hash": "7bfacfb4740e4db83f0c0b4042bd34c6",
  "upload_date": "10:54:57 05/03/2024"
}
```


Figure 10: Example of a Document Stored on MongoDB's Documents Collection

A diagram representing the entire process described in this section can be seen in Figure 11. The diagram begins with the user creating a keyword list to be used in the initial filtering process. Then, the user can choose to either crawl an online repository of their choice or consult identified recent and potentially relevant documents obtained by the RSS Feed Consumer. If the user opts to consult the RSS Feed Consumer, there are two scenarios depicted in the diagram: either no relevant documents are presented, in which case the user should start the process again, or relevant documents are presented, and the user selects them to be collected by the crawler. If the user chooses to crawl specific repositories instead of consulting relevant documents, they should input the repository links into the crawler as

shown in the diagram. After collecting documents and filtering them according to the developed keyword list, the application will generate a hash for each collected document and compare these hashes with those of documents already present in the repository to identify duplicates. Duplicated documents are excluded from further operations and will not be stored in the repository. Unique documents continue through the process, with their first 850 tokens extracted and sent to the GPT API. The API will evaluate whether the document is related to cybersecurity topics and proceed accordingly. If the document is related, GPT will categorize and provide information about it, which will be stored alongside the document's hash, upload date, and the original PDF file in the database.

Finally, in this automated collection and classification process, and solely for post-analysis of the application's results and GPT API's efficiency, the Automated Repository contains four folders where copies of all crawled documents can be stored: "Accepted", "Rejected", "Manual Deleted" and "Duplicated". If a crawled document is rejected during the keyword or GPT filtering stages, a copy of it is stored in the "Rejected" folder. If it is rejected during the hash comparation process, the document is stored in the "Duplicated" folder. On the other hand, if a document passes the filtering processes, is classified by GPT, and is stored in the database, a copy of it is placed in the "Accepted" folder. If a user later deletes a document that was previously added to the repository, the copy of the document is moved from the "Accepted" folder to the "Manual Deleted" folder. This system allows for conclusions to be drawn about acceptance and rejection rates, as well as providing insights into how many accepted documents are later identified as irrelevant and subsequently deleted from the repository.

3.3 REPOSITORY OVERVIEW

After the addition process is complete, the documents will be available in the repository. In Figure 12, the Automated Repository homepage is displayed. This page provides a brief overview of the application's primary features and serves as the entry point for new users.

Following this, Figure 13 shows the main repository page. On this page, users can find all the documents contained in the repository. The repository main page is equiped with pagination, a searchbar and filtering tools.

The interface fetches documents via a GET request to the backend's API endpoint */get-documents*. Upon receiving the request, the backend queries the database and

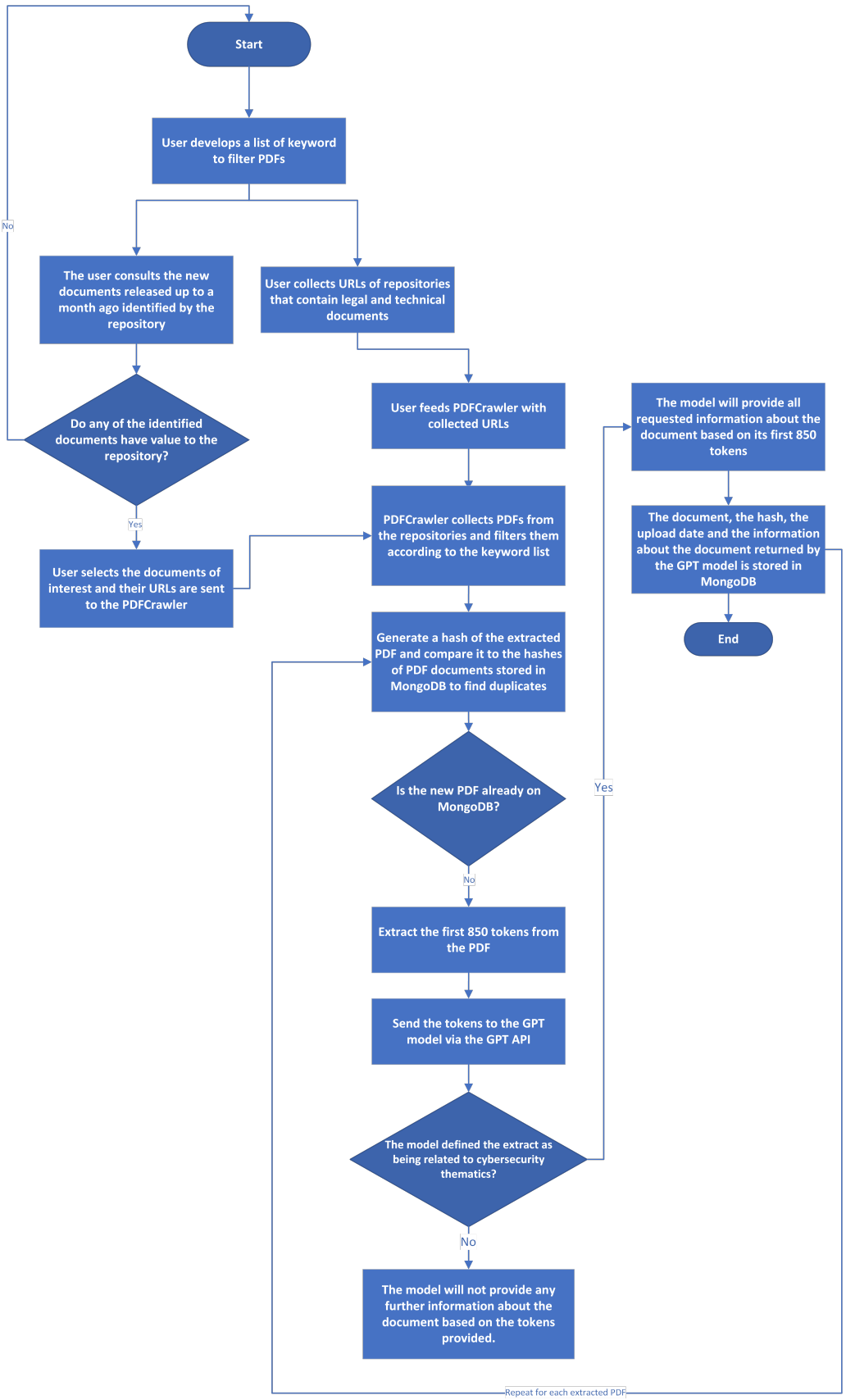


Figure 11: Add Documents to the Automated Repository Process

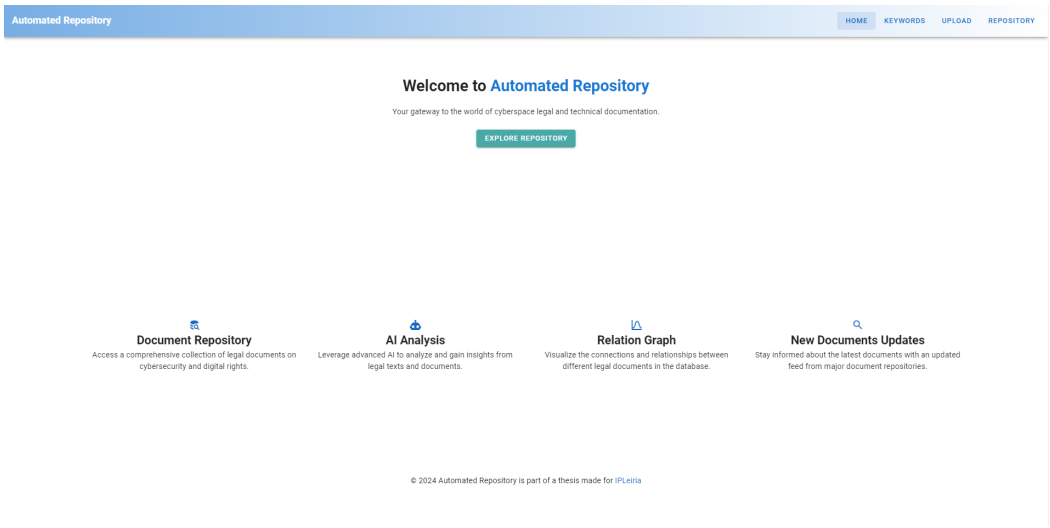


Figure 12: Automated Repository's Home Page

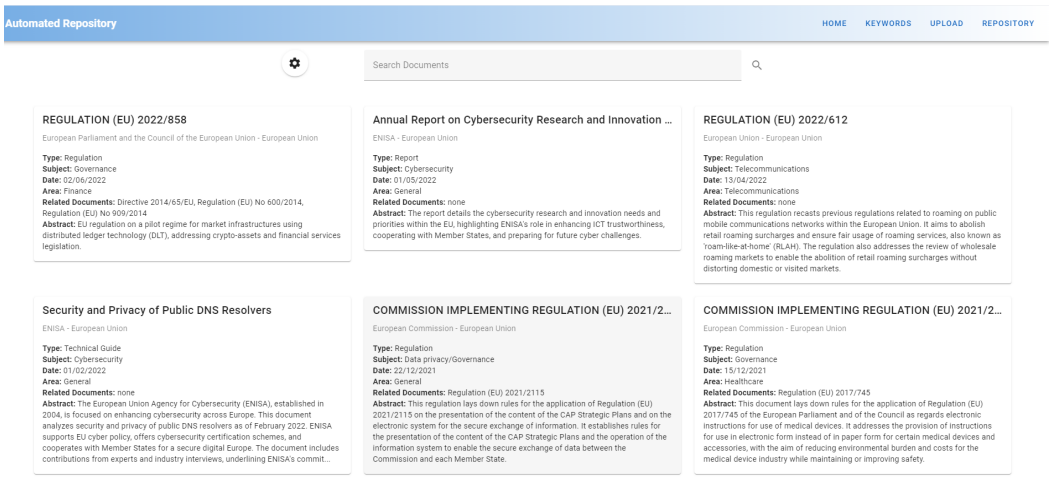


Figure 13: Repository's Main Page

returns all the documents to the frontend in a JSON's list of objects, where each object represents a document and contains all its information.

As mention previously, users have the capability to search for specific documents and apply filters based on specific areas, subjects, issuers, etc., to refine their search results. The search results can also be sorted by title and date, among other criteria.

The filtering and search logic is entirely implemented on the frontend. For filtering, custom functions have been developed specifically for this purpose. For searching, Vuetify 3's built-in capabilities are utilized.

By clicking on a specific document, users are redirected to a "Details Page" that presents all information related to that document.

When the UI detects a click on a document's card, a method is triggered that redirects the user to the URL */details/document_id* that corresponds to the "Details Page" of that same document. In this URL, *document_id* is replaced by the Identification (ID) of the document whose card was clicked. This ID corresponds to the document's ID value in the database.

The "Details Page" uses the document's ID from the URL to send a GET request to the backend at */get-document/documentId*. The backend queries the database for the specific document and returns all the document's information to the frontend in the form of a JSON object. An example of a JSON object sent from the backend to the UI can be seen on Figure 14.

```
{
  "_id": "65b7cc8081a7512f344f7aa7",
  "is_related": "yes",
  "issuer": "Presidência do Conselho de Ministros",
  "origin": "Portugal",
  "type": "Decreto-Lei",
  "subject": "Cybersecurity",
  "date": "20/12/2023",
  "area": "General",
  "title": "Decreto-Lei n.º 116/2023",
  "related_docs": [
    "Decreto-Lei n.º 150/2019",
    "Decreto-Lei n.º 65/2021"
  ],
  "abstract": "Transfers oversight and countermeasures instruction competencies
  ↪  related to ECOMPENSA to the National Cybersecurity Center (CNCS). Aims to
  ↪  simplify sanctions and enhance effectiveness.",
  "pdf_file_id": "65b7cc8081a7512f344f7aa4"
  "pdf_hash": "774ca4e4995cccd57f6235a0c3b21fdd",
  "upload_date": "10:48:55 27/03/2024"
}
```


Figure 14: JSON Object sent by the Backend to the UI contain all requested document details

The user's view of the "Details Page" can be seen in Figure 15, where an entry from our repository representing the EU's General Data Protection Regulation (GDPR) is showcased. This GDPR entry firstly displays its official title, Regulation (EU) 2016/679. Following the title, the other fields that were also generated by GPT are represented.

On the Details page, users have the ability to perform a variety of operations on the document they are currently viewing. These options are accessible through the blue bottom bar, which is also visible in Figure 15.

From left to right, a user can go back to the "Main Repository Page", manually edit all fields of the document, request GPT to regenerate a specific field (this functionality is fully described in Section 3.3.1), download the original document on which all information was based, and finally, delete the document from the repository.

If the user chooses to edit a document, they are redirected to the "Edit" page. The URL of this page follows the structure: */edit-document/document_id*.

Figure 15: GDPR Representation on the Details Page

The "Edit" page, which is represented in Figure 16, starts by loading all the information from the document in the form visible in the user interface. This information is obtained using the same process as described for the 'Details Page'. Once the information is loaded, the user can view the current values and make any necessary edits.

When all edits are completed, the user can press the save button to save the changes. This action triggers a method that sends a PUT request to the */details/document_id* endpoint. In this request, a JSON object representing the document with the newly defined field values is included in the request body. The format of this object is similar to the one already described in Figure 14.

The backend will then query the database using the document's ID and utilize MongoDB's built-in function to replace the document information associated with that ID with the updated data represented by the JSON object.

If the user chooses to download the original PDF file instead of editing it, the user will send a GET request to the back-end endpoint */get-pdf/documentId*. The backend will fetch the document from the database. Since the document is stored in chunks, MongoDB will reassemble the document and retrieve it for the backend.

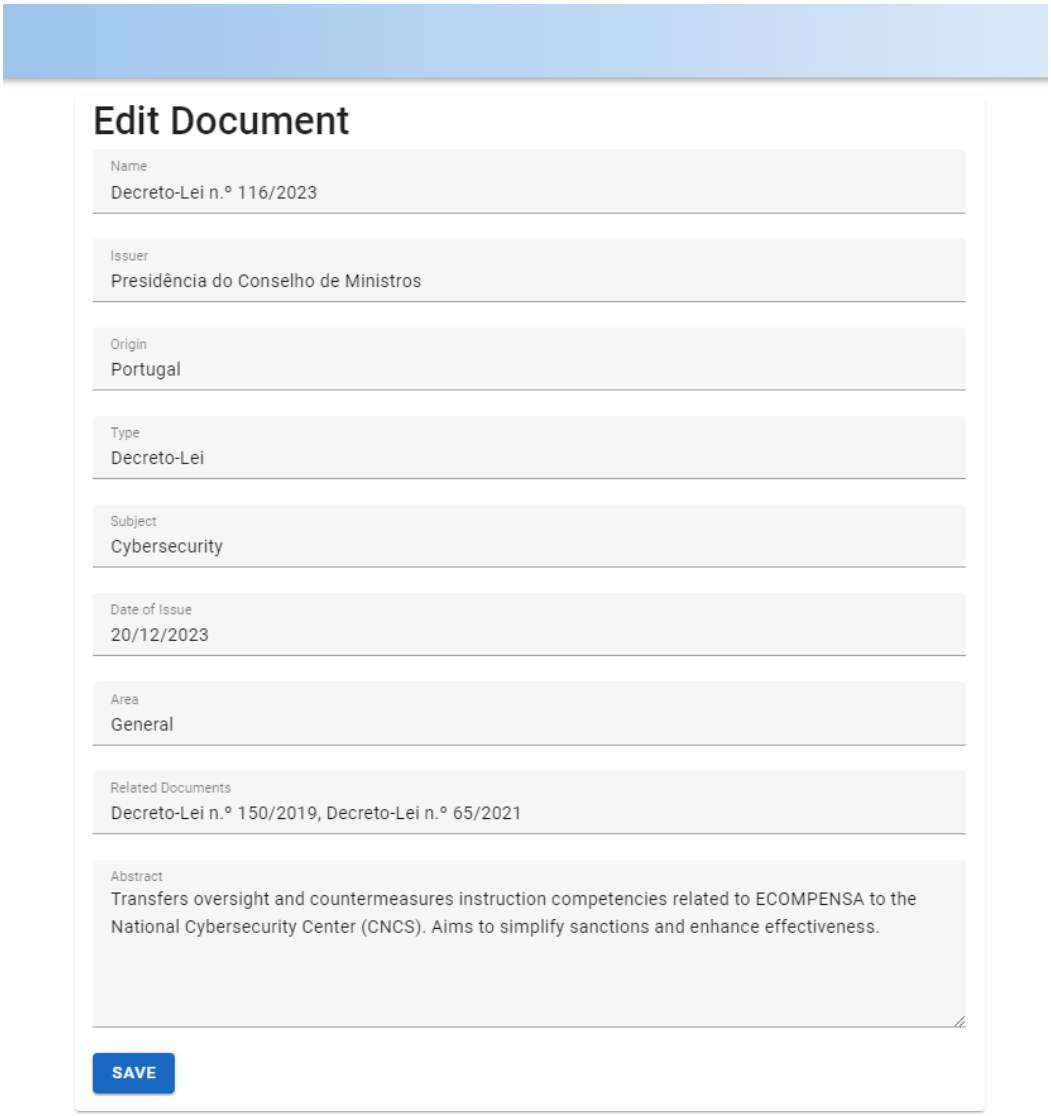


Figure 16: Manual Edition of Document's Fields

The UI will then receive the PDF and display it in a new window, where the user can view and download it.

Deleting a document involves sending a DELETE request from the UI to the backend's endpoint */delete-document/documentID*. The backend will use MongoDB's built-in methods to remove the document and all its generated information from the database. Additionally and only for post analysis porpuses, the document will be moved from the "accepted" folder to the "manually deleted" folder.

3.3.1 *Regenerate Document Fields*

A core functionality of our Automated Repository is the "Regenerate Page". Here, a user can request GPT to reanalyze any field of a certain document and define a specific creativity level for this operation. This page, that can be observed in Figure 17, will display the current values for each field. These values are obtained through a GET request to the endpoint */get-document/documentId*. It will also display a slidebar where users can define the desired creativity level for this operation.

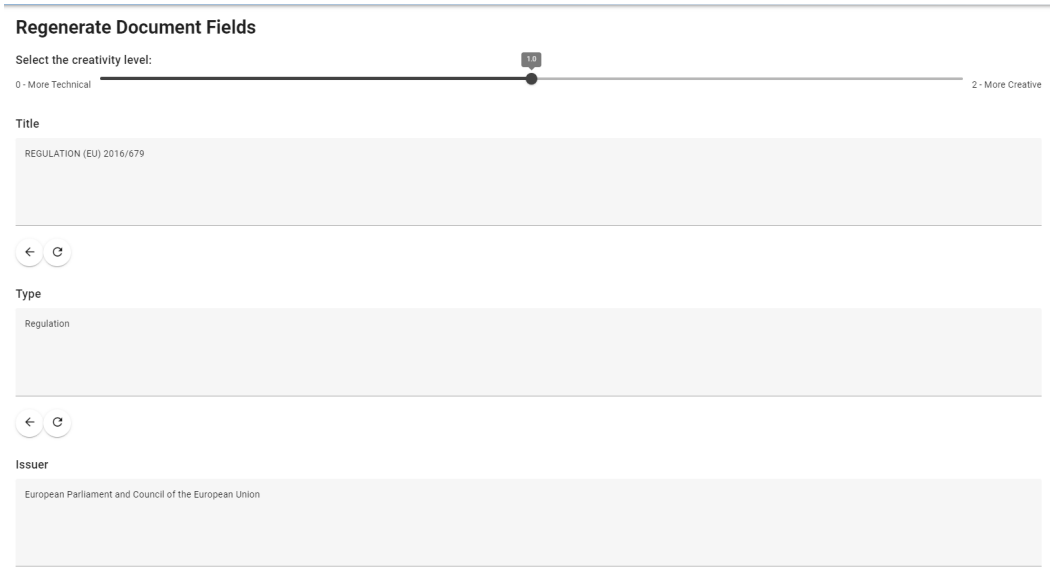


Figure 17: Regenerate Page

After selecting the field to be regenerated and the creativity level, a POST request will be sent to the backend through the */regenerateDoc* endpoint. The request structure will follow the example shown in Figure 18. As can be observed, three values are sent in this request:

- **documentID**: Is the ID that represents the document whose field is going to be regenerated
- **field**: 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**: Users can set the "Creativity Level", also known as temperature as explained in section 3.1.2, 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 18: 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 regenerated field's value will be returned by the GPT API to the backend, and sent to the UI in the form of a JSON object, following the example present on Figure :

```
{
    "abstract": "Directive 2001/29/EC, passed by the EU Parliament and Council,
    ↪ aims to harmonize copyright and related rights within the information
    ↪ society. This aligns with the EU's internal market principles, fostering
    ↪ the development and competitiveness of European industries by protecting
    ↪ intellectual property, stimulating investment, and encouraging
    ↪ creativity and innovation in IT and cultural sectors. The directive also
    ↪ supports job creation as it adapts copyright laws to technological
    ↪ advancements."
}
```

Figure 19: Object Returned By Regenerate Abstract Field

The user always has the option of reverting to the original value of the field that was first displayed when they opened the page. This can be done by selecting the back arrow below the regenerated field, as shown in Figure 17.

The "Regenerate Page" UI always stores the default values of the fields. When the user opts to revert the regeneration process, the frontend will simply replace the newly generated field's information with the default values.

To make the regenerated values permanent, the "save" button on the bottom of the page should be clicked. When pressed, the button starts a PUT request to the backend's */update-document/documentId* endpoint, in a process similar to the one described in the "Edit Page".

The final aim of the 'Regenerate Fields' functionality is to offer users a more interactive experience with GPT, allowing them to specify the particular documents, fields, and levels of creativity used in the information generation process.

3.3.2 *Relation Graph*

In addition to displaying and describing the documents, Automatic Repository includes the functionality to showcase the relationships between them.

This is achieved through a network graph, which is illustrated in Figure 20.

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

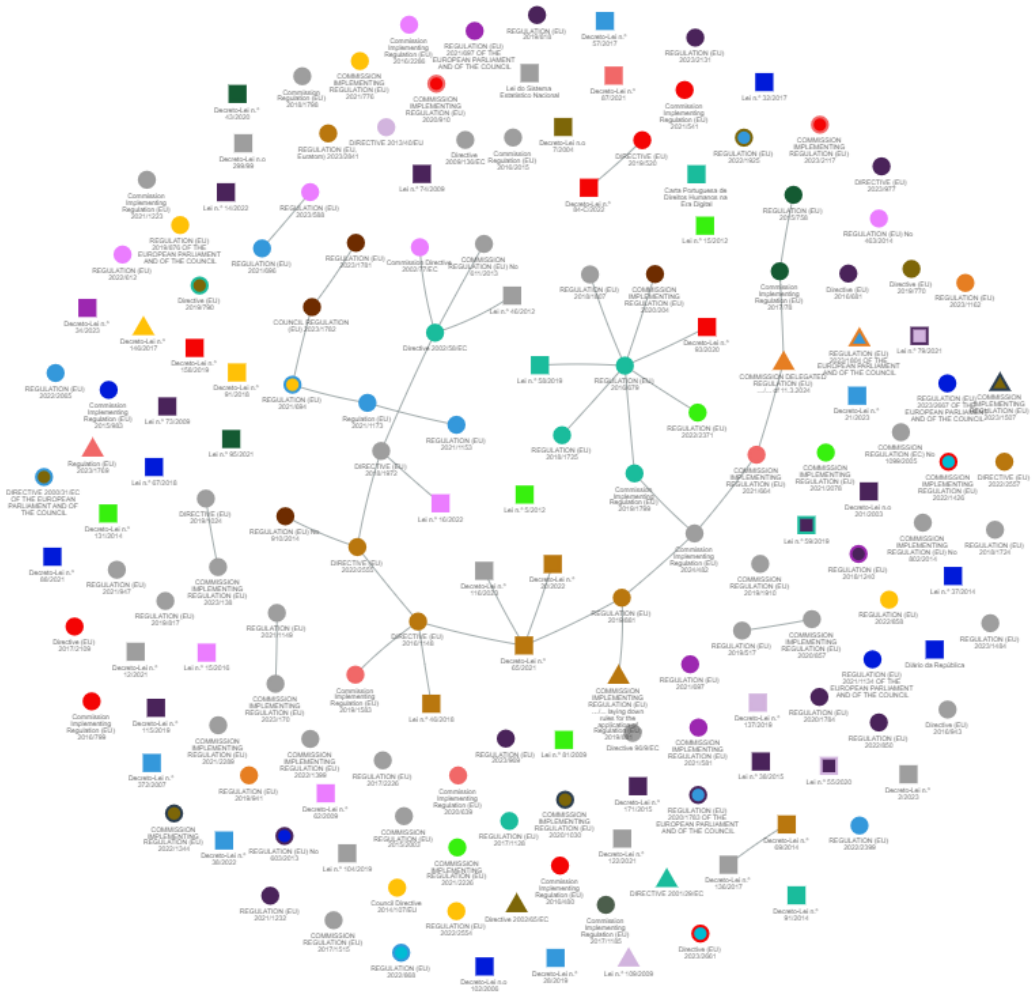


Figure 20: Relation Graph

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 our Automatic Repository, that particular relationship will not be depicted in the graph.

Figure 21 represents a segment of the relation graph, where we can closely observe that in addition to displaying relationships, each node possesses additional characteristics. Below each node is the title of the document it represents. Furthermore, it is evident that the nodes differ in color and shape.

The colors of the nodes signify the areas of the documents they represent, with each area associated with a specific color. The color that corresponds to each area is described in Table 1. The user can also view the meaning of the different colors and shapes in a caption provided in the UI.

Area	Color	Hex Color Code
Agriculture	Dry Green	#4C5E4C
Artificial Intelligence	Cyan	#00BCD4
Cybersecurity	Gold	#B9770E
Defense	Violet	#9C27B0
Digital ID	Dark Blue	#001AD9
Digital Rights	Marine Blue	#1ABC9C
Digitalization	Sky Blue	#3498DB
e-Commerce	Dark Brown	#7D6608
Electronics	Brown	#6E2C00
Emergency Services	Dark Green	#145A32
Energy	Orange	#E67E22
Finance	Yellow	#FFC107
General / Other Areas	Gray	#9E9E9E
Healthcare	Light Green	#37F00E
Justice	Indigo	#4A235A
Telecommunications	Pink	#EB7DFF
Transport	Red	#F40404

Table 1: Area´s colors on the Realtion Graph

Figure 21 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 the figure, 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.

As it was previously mentioned in Section 3.1.1 and represented in Figure 2, the frontend is responsible for generating the relation graph, as well as managing all interactions with it.

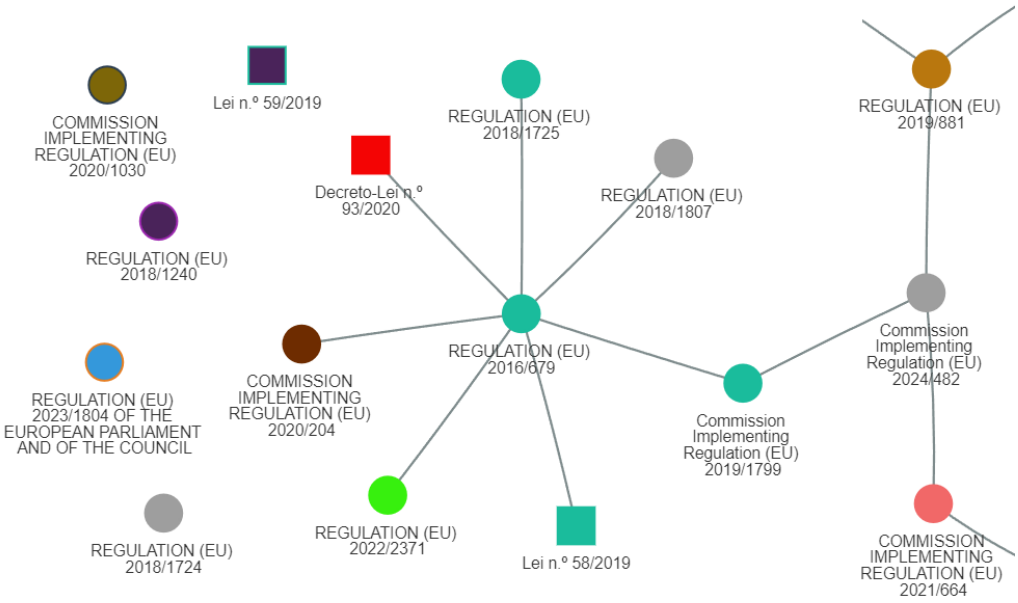


Figure 21: Relation Graph Detail

When the user first opens the page, a GET request is sent to the backend endpoint "/get-documents". This endpoint returns a list of objects, with each object representing a document and its attributes.

For each document, the frontend will extract the ID, areas, title, and origin, and will generate a unique node for each document. It will also draw relations between these nodes as previously described. All nodes and relations are drawn with the help of the vis-network library [61], as mentioned in Section 3.1.1.

Users can also click on each node. If a click is detected, the UI 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 that specific node, and the node will be highlighted.

The Relation Graph offers valuable insights to the user by illustrating how cybersecurity documents are interconnected. It also highlights certain documents that are central within the cybersecurity landscape, as they are connected to many other nodes and have a significantly higher number of relationships compared to other documents.

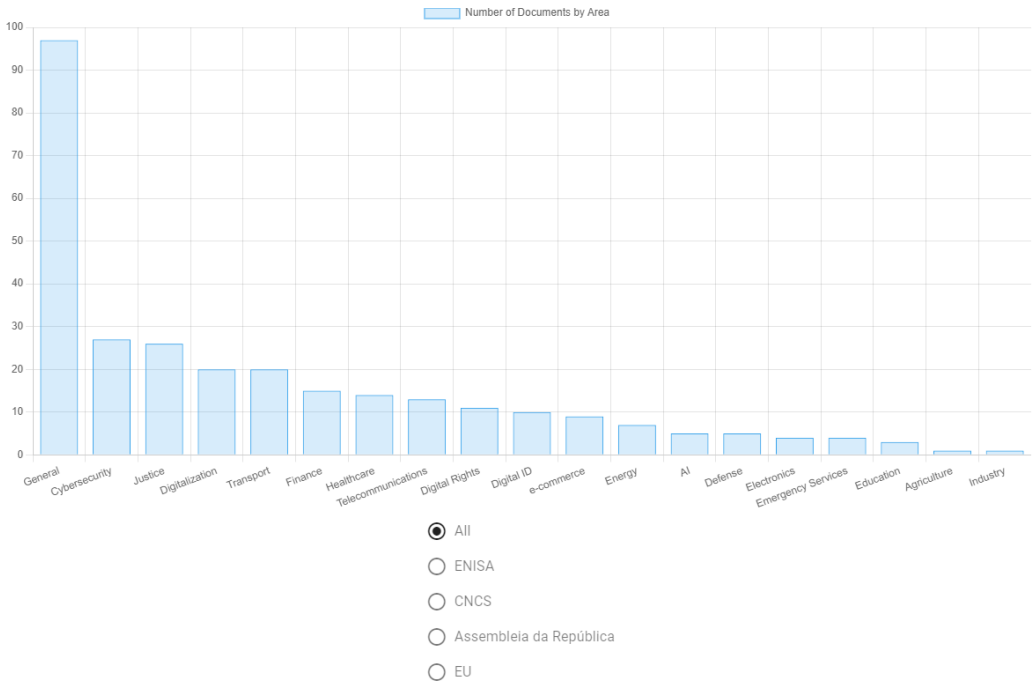


Figure 22: Number of Documents Grouped by Area

3.3.3 *Statistics*

The Automated Repository also includes a "Statistics Page". This page contains a series of data visualizations that provide users with insights into the repository's contents. The visualizations include:

- Number of documents grouped by area, from all or difrent issuers - Figure 22.
- Number of documents issued over time and grouped by different origins - Figure 23.
- Comultative and mounthly count of ducuments present in the repository - Figure 24.
- Number of documents grouped by its type (Law, Decree-Law, Report, etc) - Figure 25
- Number of documents issued per year and per area - Figure 26.

For the Statistics Page, the rendering of the graphs for information display is entirely performed by the frontend, but the information is preprocessed by the backend. For all five graphs on this page, the frontend starts by making GET requests to the backend. The endpoints vary depending on the graph, but all endpoints are called as the graphs are generated when the user opens the page. The

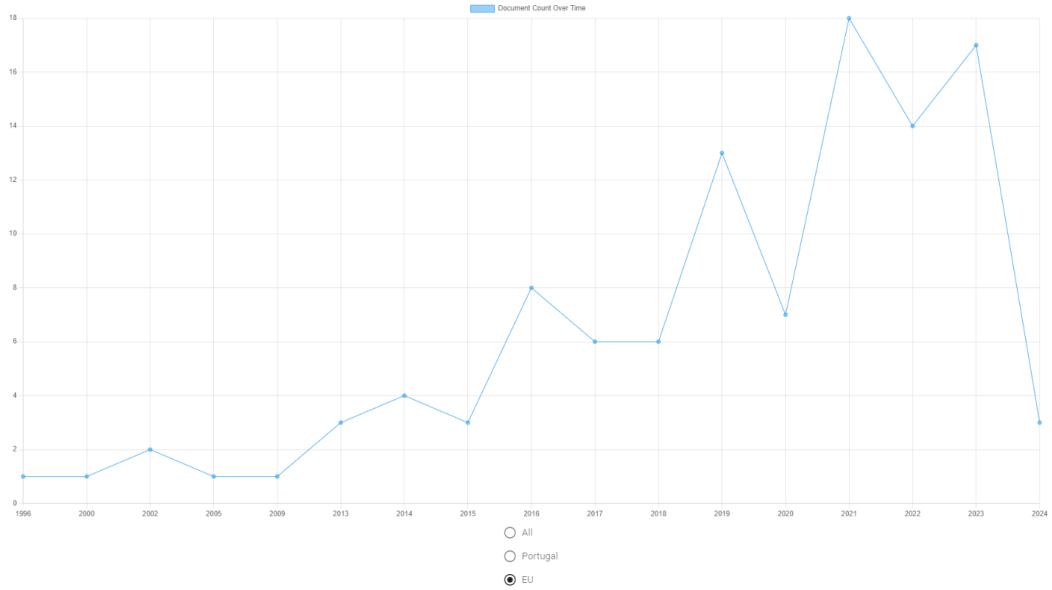


Figure 23: Documents Issued Over Time and Grouped by Origin

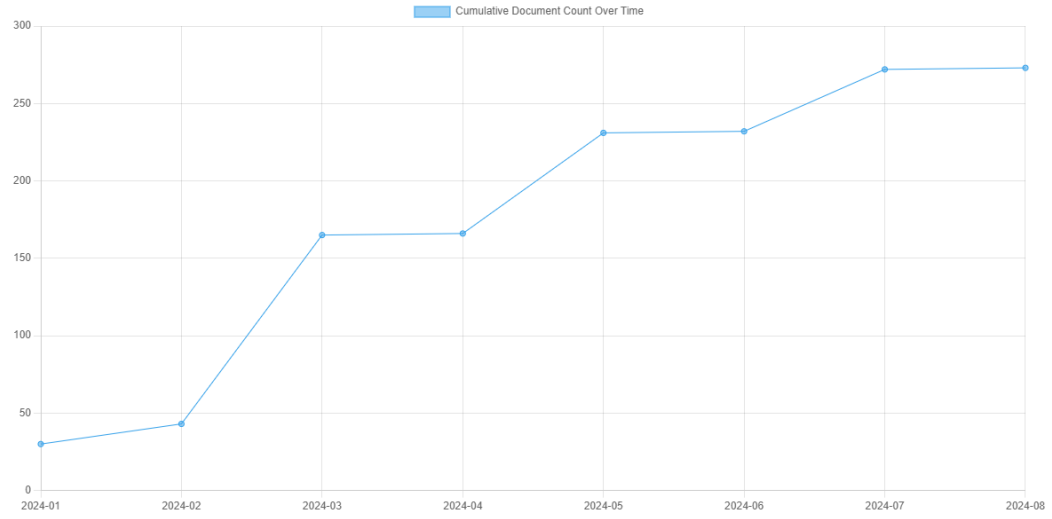


Figure 24: Comultative Count of Documents Present in the Repository

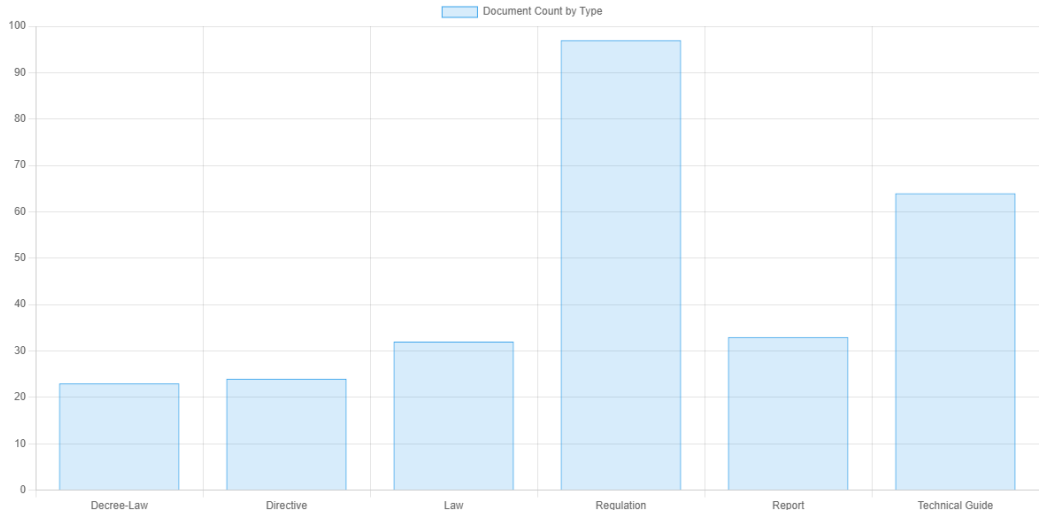


Figure 25: Number of Documents by Type

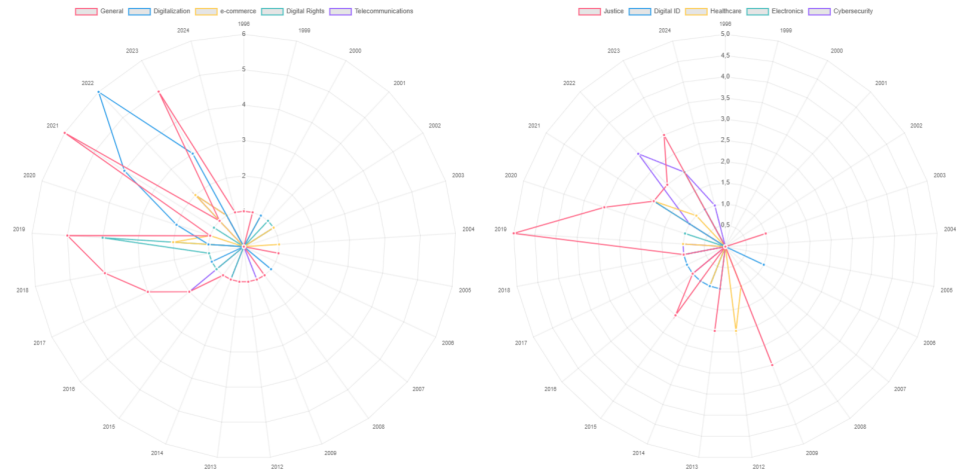


Figure 26: Documents Issued Per Year and Per Area

endpoints are */areas*, */document_counts_by_year*, */document_counts_by_month*, */document_counts_by_type* and */area_counts_by_year*.

For each endpoint call, the backend performs the necessary queries to the database to prepare the data for display by the frontend. These queries are performed using MongoDB's pipeline capabilities.

MongoDB pipelines can be seen as a sequence of steps (stages) used to process and transform data stored in a MongoDB database. Each step (stage) in the pipeline performs a specific task, such as filtering data, grouping similar items, or calculating totals. The data flows through each step in order, and the result of one step is passed to the next. This allows to efficiently analyze and manipulate data directly within the database.

There are a set of steps predefined that can be utilized in pipelines:

- **$match:** Filters documents based on a condition
- **$group:** Groups documents by a specified key and performs aggregate functions like sum, count, average, etc
- **$project:** Shapes the documents by including, excluding, or adding fields
- **$sort:** Sorts documents based on specified fields
- **$unwind:** Takes an array field from the documents and creates a separate document for each element in the array

For example, a pipeline that can be used to count documents grouped by area in Figure 27.

```
pipeline = [
        # Split the 'area' field if it contains '/'
        {"$project": {
            "areas": {
                "$split": ["$area", "/"]
            }
        }},
        # Unwind the array to handle documents with multiple areas
        {"$unwind": "$areas"},
        # Trim any leading or trailing whitespace
        {"$project": {
            "area": {"$trim": {"input": "$areas"}}
        }},
        # Group by the trimmed 'area' and count each one
        {"$group": {"_id": "$area", "count": {"$sum": 1}}}
    ]
```

Figure 27: Documents By Area Pipeline

After the pipeline processing is complete, the prepared data is sent to the frontend in JSON object format, as shown in the example in Figure 28. This data is then plotted in graphs using the Chart.JS library [62].

```
{
    "AI": 3,
    "Agriculture": 1,
    "Cybersecurity": 11,
    "Defense": 5,
    "Digital ID": 10,
    "Digital Rights": 11,
    "Digitalization": 19,
    "Electronics": 4,
    "Emergency Services": 4,
    "Energy": 4,
    "Finance": 9,
    "General": 37,
    "Healthcare": 8,
    "Justice": 25,
    "Telecommunications": 8,
    "Transport": 17,
    "e-commerce": 9
}
```

Figure 28: JSON Object Used in Graph Creation

With the statistics page, users can gain a better understanding of the repository's contents, and better comprehend the evolution of cybersecurity and cyberspace related documents over time and their distribution across different areas. This type of information is particularly useful to draw insights about the evolution of the cybersecurity related areas that have been most impacted by new regulations and legislation.

3.3.4 *New Documents Page*

Users can provide URLs for repositories or documents relevant to the repository. Additionally, "New Documents Page", which is represented in Figure 29, can notify users of the latest documents that might be of interest and can be added to the repository.

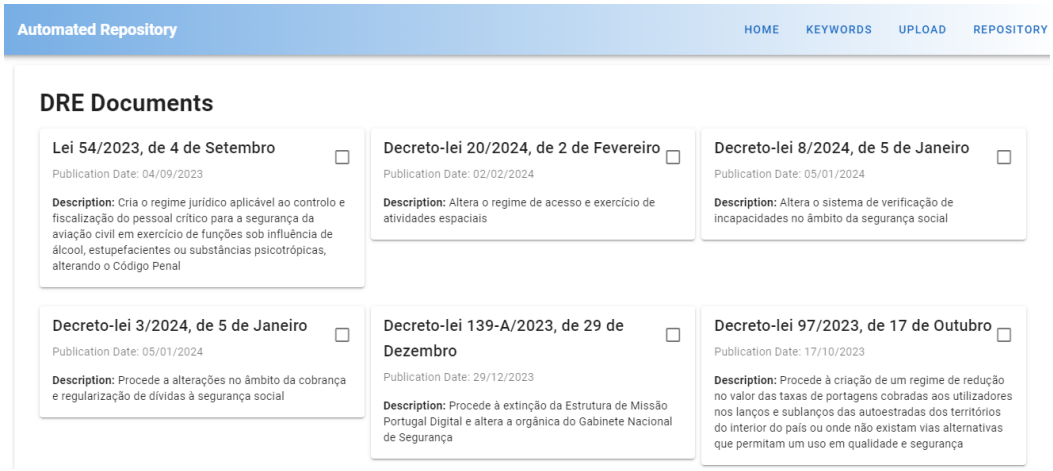


Figure 29: New Documents Page

Using the "New Documents Page", users can select individual 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, allowing the PDFCrawler to fetch and store them for future analysis using 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 3.1.5.

Each card displays the document's title, date, and a short description, helping users easily determine the relevance of the document to the automated repository.

3.4 IMPLEMENTATION SCENARIOS

Using the installation guide provided on the Automated Repository GitHub page and referred to in Section 1.3, this project was implemented in three main scenarios:

The first scenario, where the application's capabilities were fully explored, involved an installation on a Windows 11 machine with 16 GB of RAM and an 8-core processor. This setup was used to develop the results discussed in Chapter 4.

Subsequently, the Automated Repository was also tested on a Windows 11 machine with 4 GB of RAM and a 2-core processor. The same hardware specifications were also used for an Ubuntu 24.04 Desktop installation.

On all three cases, all three components (database, backend and frontend) were installed on the same machine.

Automated Repository demonstrated acceptable performance with only 4 GB of RAM, but it is still recommended to use our application on systems equipped with at least 8 GB of RAM. This is because the crawler can greatly benefit from the additional memory. This is also true for visual tools performance, especially when those need to perform operations like graph generation, which can become resource-intensive with a high volume of documents in the repository.

4

RESULTS ANALYSIS

This chapter presents an analysis of the results obtained from the development of the automated repository. It also describes some conclusions drawn from the development process and enumerates many relevant observations regarding GPT behavior.

4.1 DEVELOPMENT CHALLENGES AND OBSERVATIONS

During development and repository populating process, many relevant observations were taken. In the following subsections, all these observations are addressed.

4.1.1 *Populating the Repository*

A total of 567 documents were obtained from "Diário da República" [36], "Eur-Lex" [33], ENISA Publications [29] and CNCS repositories [18]. These repositories and entities are described in Sections 2.1 and 2.3. The documents were identified using two different methods:

- By utilizing 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 UI of the same page.

After collection, and as detailed in Section 3.2, the process of classifying documents into categories related or not related to cyberspace by GPT began.

A total of 376 documents were classified as being relevant for our Repository, while 183 were rejected by the model and 8 documents were identified as duplicated.

4.1.1.1 *Issues during Document Filtering*

During the filtering process, it was observed that GPT occasionally made errors. In particular, it incorrectly identified documents with only slight mentions of IT-related topics as relevant to the theme, suggesting a challenge in discerning the actual relevance of such mentions.

For instance, GPT identified "Lei n.º 46/2019" as relevant to our repository. However, this document pertains to legislation for the operation of private security organizations. Although it covers a range of topics, it includes minor points that refer to the proper operation of video surveillance systems and the correct way to manage such data.

This identification made by GPT is not entirely incorrect, but it is easily debatable. Having a brief section addressing data privacy does not make a document relevant to a repository intended to exclusively contain documents about cybersecurity legislation and regulation.

Another example of this issue is the decree law "Decreto-Lei nº 19/2022", which GPT also identified as relevant to the repository. This document is a complex and extensive legal text detailing numerous reforms to the organization of the Portuguese Armed Forces. Although the document briefly mentions that cyberspace should be the focus of a specific unit within the armed forces, cyberspace and related topics are not the main focus of this document by any measure.

Other error GPT made while filtering relevant and non-relevant documents was occasional but very peculiar. In the field "is_related", the model entered the value 'Yes' and proceeded to generate information for all other fields. However, when the model reached the Abstract field, GPT generated a standard abstract as if for any other document. Only in the end, and in uppercase, did it state: "THIS DOCUMENT IS NOT RELATED TO IT/CYBERSECURITY." There were only 4 cases when GPT made this mistake, on the entire universe of documents analyzed, and in all cases the document in question was in fact not related to any topic relevant to cybersecurity legislation.

The final issue observed in this filtering process arises not directly from GPT, but from poorly organized documents. Specifically, some decree-laws extracted from the "Diário da República" were problematic. These documents, intended to describe a single decree-law, instead contained a series of consecutive decree-laws. Additionally, since GPT only had access to the first 850 tokens of text, it often only saw part of a previous decree-law, not the one it was supposed to analyze. If the preceding

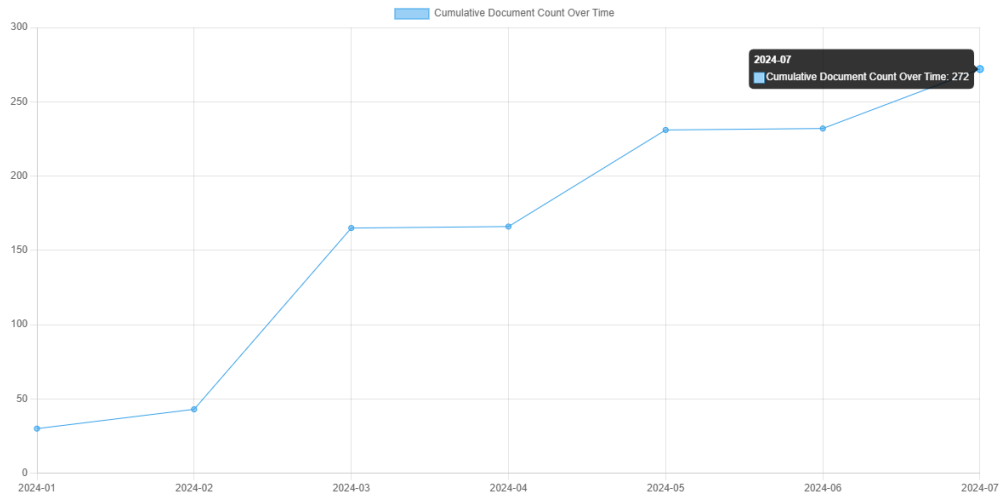


Figure 30: Number of Documents in the Repository by Month

decree-law wasn't related to cybersecurity, GPT classified the entire document as irrelevant to the Repository.

A total of 104 documents were incorrectly identified as relevant, representing about 28% of the documents deemed relevant and 18% of all documents analyzed. These percentages could have been lower due to two factors: the use of GPT-3.5 Turbo for a significant portion of the analyzes to reduce costs, since 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, as previously noted.

Currently, the repository contains 272 cybersecurity documents that users can observe using the various user interfaces and statistics present in the application.

Figure 30 contains a line graph from the "Statistics Page" of the Automated Repository that displays the evolution of the number of documents available in the application during the months. The graph also shows that the growth in the number of documents has been constant since the beginning of the application's development.

4.1.1.2 *Observations During Document Classification*

After distinguishing between relevant and non-relevant documents, GPT proceeded to classify each document as described in Section 3.2.

Depending on the attributes GPT generated for each document, different behaviors could be observed.

- **title**: Most of the titles generated by GPT matched the official document's titles, aligning with the intended behavior. However, there were instances where the titles generated were too extensive. Furthermore, the majority of titles were generated in all capital letters, which lacks visual appeal. A simple prompt refinement could easily fix this last issue.
- **Subject**: GPT effectively identified the subjects of all documents with few issues. Occasionally, it confused terms meant for categorizing areas with those meant for subjects, leading to some inaccuracies. A refinement of the prompt to restrict the terms used for defining subjects successfully resolved this issue.
- **Areas**: GPT successfully identified the document areas in straightforward cases. However, in scenarios requiring interpretation of the document or its context, GPT often labeled the area as "general".
- **Related Documents**: This attribute presented many challenges for GPT, as it only identified documents as related if those were mentioned in the first 850 tokens provided for analysis. Consequently, a significant number of documents were entered into the repository without any references in their related documents attribute.
- **Abstract**: This attribute showcased GPT's best performance. The model successfully generated short but informative abstracts that provided key insights into the document's aim, context, content, and objectives. Interestingly, several abstracts also mentioned related documents that were not identified in the "related documents" attribute.
- **Other Attributes**: GPT showed no significant issues while generating the "issuer", "origin", "type", and "date" attributes.

4.1.1.3 *API Usage and Total Monetary Cost*

In this project, four GPT models were used: GPT-3.5 Turbo, which was later replaced by GPT-4o mini, and GPT-4 Turbo, which was replaced by GPT-4o.

The total cost of using the GPT API in this project was around $4.40 or €4.03. Of this amount, $2.63 corresponded to the GPT-4 Turbo model, representing approximately 60% of the total expenditure. GPT-4 Turbo was the primary model used in the project because GPT-4o was only released on May 13th, 2024, while the development of this project began in February 2024.

A total of 636,463 tokens were sent or generated by the API, with the vast majority (around 85%) being input tokens. The model that received and generated the most tokens was GPT-4 Turbo, again because GPT-4o is a more recent model. This token consumption and generation was the result of a total of 634 API requests made to GPT.

4.2 CYBERSECURITY DOCUMENTATION LANDSCAPE

By analyzing the documents in the repository, including the documents fields generated by GPT and the data visualizations provided by the repository, it is possible to draw some conclusions about the current state of cybersecurity legislation and technical documentation in Portugal and in the EU.

First, it should be noted that cybersecurity regulation and documentation are in constant evolution. In the repository, the oldest document is Directive 96/9/EC from the EU, dating back to 1996. This directive was about the legal protection of databases, that should be done throughout authorship restrictions. Since then, cyberlaw, technical norms, and governance documentation have evolved significantly.

It is also possible to notice that the current authorities that are responsible for the regulation of cyberspace and cybersecurity both in Portugal and in the EU have a very recent history. ENISA was only created in 2004 with the Regulation (EC) No 460/2004. Since then, the regulation that found ENISA has suffered several changes, the most recent being the Regulation (EU) 2019/881 which replaced the previous regulations and that is included in our repository.

Portugal's CNCS was born in 2014 with the Decree-Law 69/2014. With the Decree-Law 136/2017, CNCS obtained its current name and gained new responsibilities. One year latter, in 2018, the law 46/2018 defined CNCS as the Portugal's national authority for cybersecurity.

The data collected in the repository also show that the EU and Portugal have been quickly developing new laws and regulations to address emerging cybersecurity challenges. In the years following 2008, a significant increase in the number of documents related to cybersecurity legislation and governance is visible. The year 2009 saw a peak of 6 documents released. From these documents, 5 issued in Portugal and 1 in the EU, the Portuguese law nº 109/2009 stands out. This law approved the cybercrime law in Portugal and established material and procedural penal provisions, as well as provisions on international cooperation in criminal matters,

relating to the domain of cybercrime and the collection of evidence in electronic format. In the same year, other laws regarding the processing and treatment of criminal data were also approved. Also in 2009 the European Union released the Directive 2009/136/EC, which emphasizes users' rights and privacy protection needs in the electronic communications sector. This directive, which is still in force, can be seen as a precursor to the GDPR that was approved in 2016.

Between 2010 and 2013 a total of 6 documents were released, 3 in Portugal and 3 in the EU. From these, 3 documents stand out:

- Portugal's law n.º 5/2012, that regulates personal data processing requirements for national health data files within the National Health Service using IT. It emphasizes data protection and privacy
- The EU's regulation n.º 611/2013, that establishes notification requirements for personal data breaches in electronic communication
- EU's directive 2013/40/EU that focuses on preventing attacks against information systems and aims to establish minimum rules concerning the definition of criminal offenses and relevant sanctions, as well as to improve cooperation between competent authorities

An increase in the number of cybersecurity legal and technical documentation was observed in 2014, when 10 documents were published in that year. This year saw the creation of CNCS as mentioned before and also the issuing of the Portuguese "Chave Móvel Digital" with Law nº 37/2014.

After 2014 and until 2020, the number of published documents showed an almost consistent growth. Between this period of time, fundamental documents were published, such as the GDPR and NIS1 Directive in 2016, and also the Portuguese law nº 46/2018 in 2018, that defined CNCS as the national authority for cybersecurity. Other notable documents issued in this time period and that are included in the repository are:

- Portugal's Law nº. 38/2015 that defines the security criteria for the system that guaranties the interoperability of the different polices IT systems and criminal databases
- EU's 2016/2286 regulation that abolished the costs for roaming services in the EU
- EU's regulations 2017/2226 and 2018/1240 that address information exchange and data registry of third-country nationals at external EU borders (Regulation 2017/2226) and the establishment of the European Travel Information and Au-

thorization System (ETIAS) (Regulation 2018/1240) that aims to strengthen and improve IT systems, data architecture, and information exchange in the area of border management, law enforcement, and counter-terrorism.

- EU's regulation 2018/1724 that establishes a single digital gateway to provide access to information, procedures, and assistance services, aiming to facilitate citizens and businesses' activities within the internal market. This regulation is a key element of the Single Market Strategy and the Digital Single Market Strategy, recognizing the role of digital technologies in transforming citizens' and businesses' lives and facilitating opportunities for innovation, growth, and jobs.
- EU's regulation 2019/517 that focuses on the implementation and functioning of the ".eu" top-level domain name
- EU's regulation 2019/881, which establishes ENISA and EU framework for cybersecurity certification

In 2020, the number of documents issued dropped significantly, reaching a low of 13 documents, which is half the number of documents issued in 2019. Some interesting cybersecurity related documents were issued during this time period marked by the COVID-19 pandemic, such as the EU regulation 2020/639 that amends the rules for the operation of unmanned aircraft systems inside the union or the Portuguese decree-law 43/2020 that establishes the Portuguese National System for Civil Emergency Planning. This plan, which follows guidelines defined by NATO, aims to ensure state functionality and security during crises. It outlines the role of the Portuguese civil emergency and protection authority in planning and coordination, defining national emergency policies in areas such as water, health, and cybersecurity.

The year 2021 was the year with the highest number of documents issued, with a total of 45 documents issued, 17 in Portugal and 28 in the EU. From these documents there is one that marked Portugal's cybersecurity history: decree-law 65/2021. This document defines new cybersecurity requirements for critical entities, assigns new responsibilities to CNCS and new guidelines for cybersecurity certifications within the national territory.

Other notable documents issued in 2021 are:

- EU's regulation 2021/696 that establishes the Union Space Program and the EU Agency for the Space Program. It aims to maintain the EU's competitive edge in space amidst emerging players and new technologies. For this end,

this document highlights the importance of digital technologies in the space sector.

- EU's regulation 2021/697 which approves the European Defense Fund in response to the changing geopolitical context and security challenges faced by the union. Cyberdefense is one of the point of interest of this document. Similar to this regulation, EU's 2021/697 regulation also addresses the EU's security and defense policy, and discusses geopolitical changes, emerging threats like cyberattacks, and the need for increased common defense efforts in the EU.
- Portugal's "Human Rights Chart on the Digital Era", that outlines the rights and responsibilities in the digital environment, emphasizing free access to the internet, promotion of equality, and elimination of barriers for people with special needs.

There were also interesting technical guidelines and reports issued in 2021. Some examples are:

- CNCS's report, "Cybersecurity in Portugal: Public Policies" provides an in-depth analysis of Portuguese public policies related to cybersecurity from both national and European perspectives. Policies, programs, and measures across various sectors are also discussed.
- The ENISA technical guide, "Cloud Security for Healthcare Services", offers several recommendations for healthcare providers and emphasizes the importance of cybersecurity in this sector.
- Also from ENISA, "EU Cybersecurity Initiatives in the Finance Sector" reviews European efforts to improve the cybersecurity posture of financial organizations in the union. It includes information on policy development, information sharing, capacity building, crisis management, awareness raising, training, standardization, and innovation.
- CNCS's "Portugal's Cybersecurity Report Society 2021" evaluates the knowledge, awareness, and relationship of society with cybersecurity topics. It also discusses other areas such as the relationship of public administration with cybersecurity (including plans, policies, procedures, etc.), the teaching of cybersecurity in education, and more.
- ENISA's "PSIRT Expertise and Capabilities Development" provides a study of Product Security Incident Response Teams (PSIRTs) capabilities in the health and energy sectors. This document also contains recommendations for these entities. Similarly, ENISA's "CSIRT Capabilities in the Healthcare

Sector" focuses on the capabilities of CSIRTs in the healthcare sector and provides recommendations to improve these teams' capabilities.

Although the number of documents issued in 2022 was slightly lower compared to the previous year, the year 2022 saw the publication of documents that marked the European cybersecurity landscape: The NIS2 and DORA. Regulation 2022/2554, also known as DORA, focuses on enhancing the digital operational resilience of the financial sector and introduces several new requirements for the implementation of cybersecurity measures and policies. NIS2 or Directive 2022/2555, is an improvement of NIS1 in order to respond to the rapid development of cyber threats. Although not as demanding as DORA, NIS2 does require all critical service providers, as well as new entities covered by this regulation, to implement a set of cybersecurity measures and policies, that do guarantee a good level of cybersecurity and resilience against threat actors.

Other notable document issued in 2022 is EU's directive 2022/2557, that ends up by being a junction between DORA and NIS2. This directive requires critical service providers to implement a set of cybersecurity strategies that were also recommended in DORA. Another interesting regulation is the EU's regulation 2022/1426, that describes the technical specifications for the approval of automated driving systems in fully automated vehicles.

Several interesting reports and guidelines were also published that year. For example, "ENISA 5G Security Controls Matrix" provides cybersecurity recommendations for 5G networks, and 'Security and Privacy of Public DNS Resolvers' reviews the cybersecurity measures implemented by the largest domain name system (DNS) providers in the market. Furthermore, in 2022, CNCS published "Portugal's Cybersecurity Report - Economy", which reviews the national cybersecurity landscape from an economic perspective, analyzing the impact of cybersecurity on various types of businesses, enterprises, sectors, and demographics.

In 2023, 42 documents were issued, 12 in Portugal and 30 in the EU. EU's Directive 2023/977 and Regulation 2023/2131, were approved, with the objective to facilitate the exchange of digital information for cases of transnational crime (2023/997) and terrorism (20223/2131). Also, EU's regulation 2023/1781 was issued that year, which focuses on Establishing measures to strengthen Europe's semiconductor ecosystem, reduce dependencies, enhance digital sovereignty, and improve supply chain resilience. Finally, in that year, regulation 2023/138 was approved. This regulation specifies a list of specific high-value datasets and the arrangements for their publication and

reuse. It aims to ensure that public data of high socio-economic potential is made available for re-use with minimal legal and technical restrictions and free of charge.

On the domain of reports and technical guidelines, 2023 was marked by the publishing of some important documents. Examples are as follows:

- ENISA's "A Governance Framework for National Cybersecurity Strategies" provides a guide to assist EU countries in developing and improving their national cybersecurity strategies. This guide aims to improve the cyberspace security of target countries and ensure compliance with the requirements of NIS1 and NIS2.
- ENISA also published "Developing National Vulnerability Programs". This guide begins by addressing the perspective of European industries on Coordinated Vulnerability Disclosure policies and challenges. It then provides several recommendations to improve vulnerability management, such as knowledge sharing, the use of open-source software to enhance collaboration, and investment in automation and bug bounty tools.
- CNCS's "Portugal's Cybersecurity Report - Risks and Conflicts 2023" provides valuable insights into the threats that affected Portuguese cyberspace up to 2023. The report details issues such as the most reported cybercrimes, the most detected and attempted types of cyberattacks, the most affected sectors, and much more.

In the current year 2024, 25 documents have been issued so far. The regulation issued in this present year that stands out the most is EU's regulation 2024/1689, also known as the AI act. The main objective of this regulation is to ensure that AI systems respect human rights and freedoms, comply with democratic values and respects users privacy. The regulation also intends to prevent legal fragmentation, ensure free movement of AI-based goods and services, support innovation, and make the EU a leader in AI adoption.

In addition to the AI act, other legal documents related to cybersecurity have been issued so far in 2024. Among these is EU Regulation 2024/482, which specifies the rules for the application of the "European Common Criteria-based Cybersecurity Certification Scheme" in accordance with the European cybersecurity certification framework. Regulation (EU) 2024/1366, also from 2024, is another important regulation that imposes new cybersecurity requirements and recommendations for the entities that manage electricity flows across the union.

Several interesting technical guidelines and reports have been published in this year. Examples include ENISA's guide "Best Practices for Cyber Crisis Management" which provides recommendations for EU states to prepare for and manage potential cyber crises, and CNCS's report "Cybersecurity in Portugal: Risks and Conflicts - 2024", which offers updated insights on this theme up to the end of the last year, 2023.

When it comes to the analysis of documents issued by date and ordered by area, some interesting observations can also be made.

Documents specifically aimed at cybersecurity measures, policies, and guidelines were published significantly between 2021 and 2024.

In 2021, amid the COVID-19 pandemic, a peak of documents related to the area "General" can be seen. Also in that year, 4 documents related to "Digitalization" of services were issued, a value only passed by the 6 documents related to the same area issued in the next year of 2022. These observations can be related to the lockdowns during the pandemic, that forced people in general to use online and digital services in a scale never seen before. Similarly, the publication of documents targeting the healthcare sector saw an all-time high in 2021, a year when the healthcare sector, as happened in 2020, received the most attention across the globe.

Between 2021 and 2022, the number of documents issued related to "Defense" applied to cyberspace and to "Cybersecurity" also reached an all-time high. This can also be related to the geopolitical changes and tensions observed during that period, and also to the ramping increase in cyberattacks and cyber threats observed.

Finally, between 2022 and 2024, the first set of documents related to AI were issued, a sign of the emergence of AI and a signal of what is to come in the future. Furthermore, in this period of time, the documents related to energy increased significantly, perhaps influenced by the energy crisis observed in Europe during this period.

The ramifications represented on the relation graph can also reveal interesting information about the landscape of cybersecurity documentation. Here, it is possible to notice that, the document that have influenced the most other laws, directives, regulations and decree-laws, is the GDPR. This document, issued in 2016, is directly related to 10 other documents. This is a clear sign of the importance of GDPR in the cybersecurity legislation and governance landscape.

In Portugal, the document that has influenced the most other laws, directives, regulations, and decree-laws is the decree-law 65/2021. This document, issued in

2021, is directly related to 4 other documents, which proves the impact that this decree-law had in the Portuguese cybersecurity landscape. By observing the graph it is also evident the importance of NIS1 and NIS2 which were the base for many more documents issued in the EU and Portugal.

5

CONCLUSION

During the classification process, it was noted that GPT sometimes made mistakes. Specifically, 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 important to highlight that GPT demonstrated a strong ability to produce comprehensive and accurate document summaries. It was also able to consistently extract detailed and valuable information from hundreds of documents in a highly useful manner. This proficiency often allowed the correction of errors that were made in other aspects of document classification.

While AI, 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 part of the documents identified as relevant and for certain details provided about each document, using GPT significantly saved time and effort. Additionally, the information given by the model about each document, and in particular the abstracts generated by GPT, greatly facilitated the manual refinement process.

With these observations in mind, it is worth concluding that developing 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 leverages GPT for document analysis, complemented by human moderation and corrections, and incorporates visualization tools, schemes, and diagrams. These combined functionalities are valuable across all fields that typically handle extensive documentation, including the area where this study was conducted, cybersecurity, as well as law, healthcare, natural sciences, engineering, social sciences, and many other sectors. They are particularly beneficial for professionals conducting document searches and analyses across diverse and complex repositories. The key functionalities of our Automated Repository that can be highly useful in other areas and sectors are:

- **AI-driven document selection**: GPT saved tremendous effort in detecting documents not of interest to the area of study. While many repositories permit advanced query generation, these queries require learning and expertise from users. Moreover, keyword-based queries often return documents that reference the keyword but are out of the user's context of interest. GPT selection proved to be a great complementary filter, saving much effort in selecting documents of interest.
- **Extracting basic document information**: 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, which allowed for a basic understanding of the content and objectives of 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 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 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 maintain 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 for 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, imposing new conditions on text formatting, 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 careful observation and analysis of the results. This makes the process accessible and manageable, enabling continuous improvement through iterative adjustments and evaluations.

The final product of automated repositories should always offer a more accessible and informative perspective on the area of study. Professionals and users should be able to gain insights simply by browsing the repository and consulting the provided visualization tools. The discussions provided in 4.2 should be replicable in any area where this model of repository is applied. To this end, and also as part of future work, it is necessary to equip the developed Automated Repository with the tools that could allow users to quickly and easily define new parameters or fields for document filtration and classification, as well as remove existing ones. Additionally, the ability to add new statistics, graphs, and even new insights should be incorporated to make the application more flexible and adaptable to other areas and sectors beyond cybersecurity.

BIBLIOGRAPHY

[1] Luciano Floridi and Josh Cowls. «A unified framework of five principles for AI in society». In: *Machine learning and the city: Applications in architecture and urban design* (2022), pp. 535–545.

[2] Samuel Chng et al. «Hacker types, motivations and strategies: A comprehensive framework». In: *Computers in Human Behavior Reports* 5 (2022), p. 100167.

[3] Chong Wang, Nan Zhang, and Cong Wang. «Managing privacy in the digital economy». In: *Fundamental Research* 1.5 (2021), pp. 543–551.

[4] Sara Quach et al. «Digital technologies: Tensions in privacy and data». In: *Journal of the Academy of Marketing Science* 50.6 (2022), pp. 1299–1323.

[5] Ida Lindgren et al. «Close encounters of the digital kind: A research agenda for the digitalization of public services». In: *Government information quarterly* 36.3 (2019), pp. 427–436.

[6] Muhammad Usman et al. «Compliance requirements in large-scale software development: An industrial case study». In: *Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21*. Springer. 2020, pp. 385–401.

[7] Mirosław Kutyłowski, Anna Lauks-Dutka, and Moti Yung. «Gdpr–challenges for reconciling legal rules with technical reality». In: *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*. Springer. 2020, pp. 736–755.

[8] *GPT-3.5 Turbo*. [Online; accessed 1. Jul. 2024]. OpenAI, July 2024. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[9] *GPT-4o*. [Online; accessed 1. Jul. 2024]. OpenAI, July 2024. URL: https://platform.openai.com/docs/models/gpt-4o.

[10] Michael Geist. «Cyberlaw 2.0». In: *BCL Rev.* 44 (2002), p. 323.

[11] *ISO/IEC 27000:2018*. [Online; accessed 23. Jun. 2024]. June 2024. URL: https://www.iso.org/standard/73906.html.

[12] *National Institute of Standards and Technology | NIST*. [Online; accessed 23. Jun. 2024]. June 2024. URL: https://www.nist.gov.

[13] *Assembleia da República*. [Online; accessed 6. Apr. 2024]. Assembleia da Republica, Apr. 2024. URL: https://www.parlamento.pt.

[14] *Governo de Portugal*. [Online; accessed 6. Apr. 2024]. República Portuguesa, Apr. 2024. URL: https://www.portugal.gov.pt/pt/gc24/primeiro-ministro.

[15] *European Council*. [Online; accessed 6. Apr. 2024]. European Council, Apr. 2024. URL: https://www.consilium.europa.eu/en/european-council.

[16] *European Commission, official website*. [Online; accessed 6. Apr. 2024]. European Commission, Apr. 2024. URL: https://commission.europa.eu/index_en.

[17] *European Parliament*. [Online; accessed 6. Apr. 2024]. European Parliament, Apr. 2024. URL: https://www.europarl.europa.eu/portal/en.

[18] *CNCS - Centro Nacional de Cibersegurança*. [Online; accessed 6. Apr. 2024]. Centro Nacional de Cibersegurança, Apr. 2024. URL: https://www.cncs.gov.pt.

[19] *ENISA*. [Online; accessed 6. Apr. 2024]. ENISA, Aug. 2021. URL: https://www.enisa.europa.eu.

[20] *ISO - International Organization for Standardization*. [Online; accessed 6. Apr. 2024]. ISO, Apr. 2024. URL: https://www.iso.org/home.html.

[21] *Official PCI Security Standards Council Site*. [Online; accessed 18. Sep. 2024]. Sept. 2024. URL: https://www.pcisecuritystandards.org.

[22] *Official PCI Security Standards Council Site*. [Online; accessed 6. Apr. 2024]. PCI Security Standards Council, Apr. 2024. URL: https://www.pcisecuritystandards.org.

[23] *No Direito português, qual a diferença entre uma lei, um decreto-lei e uma portaria?* [Online; accessed 5. Feb. 2024]. Feb. 2024. URL: https://ffms.pt/pt-pt/direitos-e-deveres/no-direito-portugues-qual-diferenca-entre-uma-lei-um-decreto-lei-e-uma-portaria.

[24] Infopédia. *regulamento - Infopédia*. [Online; accessed 5. Feb. 2024]. Feb. 2024. URL: https://www.infopedia.pt/apoio/artigos/$regulamento.

[25] *Tipos de legislação | União Europeia*. [Online; accessed 6. Feb. 2024]. Feb. 2024. URL: https://european-union.europa.eu/institutions-law-budget/law/types-legislation_pt.

[26] Contribuidores dos projetos da Wikimedia. *Norma técnica – Wikipédia, a enciclopédia livre*. [Online; accessed 6. Feb. 2024]. June 2023. URL: https://pt.wikipedia.org/w/index.php?title=Norma_t%C3%A9cnica&oldid=66145026.

[27] | *StandICT.eu 2026*. [Online; accessed 6. Apr. 2024]. European Standardization Observatory, Apr. 2024. URL: https://standict.eu.

[28] *Cyber Policy Portal*. [Online; accessed 6. Apr. 2024]. United Nations, Apr. 2024. URL: https://cyberpolicyportal.org.

[29] *Publications*. [Online; accessed 6. Apr. 2024]. European Union Agency for Cybersecurity, Apr. 2024. URL: https://www.enisa.europa.eu/publications#c3=2014&c3=2024&c3=false&c5=publicationDate&reversed=on&b_start=0.

[30] *National Cyber Security Strategies - Interactive Map*. [Online; accessed 6. Apr. 2024]. European Union Agency for Cybersecurity, Oct. 2021. URL: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/national-cyber-security-strategies-interactive-map.

[31] *Country Wiki - Octopus Cybercrime Community - www.coe.int*. [Online; accessed 6. Apr. 2024]. Council of Europe, Apr. 2024. URL: https://www.coe.int/en/web/octopus/country-wiki.

[32] *DataGuidance*. [Online; accessed 6. Apr. 2024]. DataGuidance, Apr. 2024. URL: https://www.dataguidance.com.

[33] *EU law - EUR-Lex*. [Online; accessed 6. Apr. 2024]. European Union, Apr. 2024. URL: https://eur-lex.europa.eu/homepage.html?locale=en.

[34] *CNCS - Observatório de Cibersegurança*. [Online; accessed 6. Apr. 2024]. Centro Nacional de Cibersegurança, Mar. 2024. URL: https://www.cncs.gov.pt/pt/observatorio.

[35] *CNCS - Quadro Nacional*. [Online; accessed 6. Apr. 2024]. Centro Nacional de Cibersegurança, Mar. 2024. URL: https://www.cncs.gov.pt/pt/quadro-nacional.

[36] *Diário da República*. [Online; accessed 6. Apr. 2024]. República Portuguesa, Apr. 2024. URL: https://diariodarepublica.pt/dr/home.

[37] [Online; accessed 27. Apr. 2024]. European Union, July 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L1148.

[38] *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. [Online; accessed 27. Apr. 2024]. European Union, Apr. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[39] *Art. 33 GDPR – Notification of a personal data breach to the supervisory authority - General Data Protection Regulation (GDPR)*. [Online; accessed 27. Apr. 2024]. intersoft consulting, Mar. 2018. URL: https://gdpr-info.eu/art-33-gdpr.

[40] *Art. 34 GDPR – Communication of a personal data breach to the data subject - General Data Protection Regulation (GDPR)*. [Online; accessed 27. Apr. 2024]. intersoft consulting, Aug. 2016. URL: https://gdpr-info.eu/art-34-gdpr.

[41] *Guidelines 2022/09 on Personal Data Breach Notification under GDPR*. [Online; accessed 27. Apr. 2024]. European Data Protection Board, Apr. 2023. URL: https://www.edpb.europa.eu/system/files/2023-04/edpb_guidelines_202209_personal_data_breach_notification_v2.0_en.pdf.

[42] *International data transfers | European Data Protection Board*. [Online; accessed 27. Apr. 2024]. European Data Protection Board, Apr. 2024. URL: https://www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en.

[43] *The EDPB: Guaranteeing the same rights for all*. [Online; accessed 27. Apr. 2024]. European Data Protection Board, June 2021. URL: https://www.edpb.europa.eu/system/files/2021-06/2020_06_22_one-stop-shop_leaflet_en.pdf.

[44] *Fines / Penalties - General Data Protection Regulation (GDPR)*. [Online; accessed 27. Apr. 2024]. intersoft consulting, Oct. 2021. URL: https://gdpr-info.eu/issues/fines-penalties.

[45] [Online; accessed 27. Apr. 2024]. European Union, Dec. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2555.

[46] *REGULATION (EU) 2022/2554 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. [Online; accessed 6. Aug. 2024]. European Union, Aug. 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2554.

[47] *Lei n.º 46/2018*. [Online; accessed 8. Jul. 2024]. Assembleia da Republica, 20918. URL: https://diariodarepublica.pt/dr/detalhe/lei/46-2018-116029384.

[48] Assembleia da Republica. *Decreto-Lei n.º 65/2021*. [Online; accessed 8. Jul. 2024]. July 2021. URL: https://diariodarepublica.pt/dr/detalhe/decreto-lei/65-2021-168697988.

[49] Trupti G Ghumade and RA Deshmukh. «A document classification using NLP and recurrent neural network». In: *Int. J. Eng. Adv. Technol* 8.6 (2019), pp. 632–636.

[50] Marco Cascella et al. «The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives». In: *Journal of Medical Systems* 48 (Feb. 2024). DOI: 10.1007/s10916-024-02045-3.

[51] Kaiz Merchant and Yash Pande. «NLP Based Latent Semantic Analysis for Legal Text Summarization». In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2018, pp. 1803–1807. DOI: 10.1109/ICACCI.2018.8554831.

[52] Degaga Wolde Feyisa et al. «The future of document indexing: GPT and Donut revolutionize table of content processing». In: *arXiv preprint arXiv:2403.07553* (2024).

[53] Abdullahi Saka et al. «GPT models in construction industry: Opportunities, limitations, and a use case validation». In: *Developments in the Built Environment* (2023), p. 100300.

[54] Jaromir Savelka et al. «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*. ITiCSE 2023. ACM, June 2023. DOI: 10.1145/3587102.3588792. URL: http://dx.doi.org/10.1145/3587102.3588792.

[55] Sengjie Liu and Christopher G. Healey. *Abstractive Summarization of Large Document Collections Using GPT*. 2023. arXiv: 2310.05690.

[56] Fatma Aladağ. «The Potential of GPT in Ottoman Studies: Computational Analysis of Evliya Çelebi's Travelogue with NLP and Text Mining and Digital Edition with TEI». In: *CULTURE* 5 (2023), p. 7.

[57] *MongoDB: The Developer Data Platform*. [Online; accessed 31. May 2024]. MongoDB, May 2024. URL: https://www.mongodb.com.

[58] *OpenAI API*. [Online; accessed 31. May 2024]. OpenAI, May 2024. URL: https://openai.com/api.

[59] *Welcome to Flask — Flask Documentation (3.0.x)*. [Online; accessed 31. May 2024]. flask, Apr. 2024. URL: https://flask.palletsprojects.com/en/3.0.x.

[60] *Vuetify — A Vue Component Framework*. [Online; accessed 31. May 2024]. Vuetify, May 2024. URL: https://vuetifyjs.com/en/#installation.

[61] *vis-network*. [Online; accessed 21. May 2024]. visjs, May 2024. URL: https://github.com/visjs/vis-network.

[62] *Chart.js*. [Online; accessed 21. May 2024]. chart.js, May 2024. URL: https://www.chartjs.org.

[63] *ChatGPT*. [Online; accessed 31. May 2024]. May 2024. URL: https://openai.com/chatgpt.

[64] *OpenAI*. [Online; accessed 31. May 2024]. OpenAI, May 2024. URL: https://openai.com.

[65] *OpenAI Models*. [Online; accessed 31. May 2024]. May 2024. URL: https://platform.openai.com/docs/models.

[66] *Pricing*. [Online; accessed 15. Feb. 2024]. Feb. 2024. URL: https://openai.com/pricing.

[67] *Pricing*. [Online; accessed 9. Aug. 2024]. OpenAI, Aug. 2024. URL: https://openai.com/api/pricing.

[68] SimFin. *pdf-crawler*. [Online; accessed 15. Feb. 2024]. Feb. 2024. URL: https://github.com/SimFin/pdf-crawler/tree/master.

[69] *Diários da República - Procura de Documentos*. [Online; accessed 28. May 2024]. May 2024. URL: https://dre.tretas.org.

[70] openai. *tiktoken*. [Online; accessed 16. Feb. 2024]. Feb. 2024. URL: https://github.com/openai/tiktoken.