TÉCNICO LISBOA

A Mapping Tool for Normative Requirements

João Cardoso Pinho da Cruz

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Miguel Leitão Bignolas Mira da Silva
Prof. Rúben Filipe de Sousa Pereira

Examination Committee

Chairperson: Prof. Paolo Romano
Supervisor: Prof. Miguel Leitão Bignolas Mira da Silva
Member of the Committee: Doutor Rafael Saraiva de Almeida

June 2023

Acknowledgments

Numerous are the people I wish to thank in this section. All of my family, who never stopped believing in me, from my parents to my grandmother, my brother, my cousins and aunt. Thank you for your overwhelming understanding and thank you for supporting me and giving me the opportunity to finish my Master's degree.

The work presented in this document is as much yours, as it is mine. I would never have gotten the chance to write these pages, were it not for you... and for that, I will always be grateful.

I would also like to thank all of my friends for their continued support. We have shared some good times and I hope we will continue to do. Not all of life is work and you've helped me maintain course in rough times. Thank you.

Last but not least, I would like to thank all of the teachers and professors that have given me all of the theoretical and practical knowledge I required to finish this document, and my degree. Especially, my dissertation supervisors, Prof. Miguel Mira da Silva and Prof. Rúben Pereira for the knowledge they shared, for their guidance and all of the insight given to me across these two semesters. Additionally, I thank my fellow student and researcher, André Fernandes, for his support in the research process.

To each and every one of you – Thank you.

Abstract

When addressing cybersecurity concerns, organizations often resort to implementing compliance mechanisms such as standards, internal policy and legislation. However, blind implementation of these mechanisms can lead to duplicated efforts when requirements overlap. To overcome these challenges, organizations can undertake mapping or integration studies to uncover commonalities and distinctions among the implemented mechanisms. However, these studies are large and complex, bringing to the tables challenges of their own. In our research, we have employed the Design Science Research Methodology and conducted a Systematic Literature Review (SLR) to identify the key research problems when carrying out the mapping and integration processes. We then design, propose, demonstrate, and evaluate a solution for the the primary issue we uncovered during the SLR, which was the time complexity associated with the mapping process. The focus of this research is to enhance the mapping process by introducing a software tool specifically designed to expedite the comparison phase between the requirements of two standards. This tool automates the comparison of requirements between two documents and produces a table with the most likely requirement overlaps between them. The tool was evaluated by gathering two mapping tables created manually, one deriving from the literature and another created by us, which serve as the baseline for comparison. We compare the results generated by our tool against the base mapping tables to assess, through various metrics we proposed, how precise the results are.

Keywords

Security; Compliance; Design Science Research Methodology; Systematic Literature Review; Standards; Artificial Intelligence; NLP

Resumo

Ao lidar com preocupações de ciber-segurança, as organizações frequentemente optam por implementar mecanismos de conformidade, tais como políticas internas, normas ou legislação. No entanto, a implementação indiscriminada desses mecanismos pode resultar em trabalho duplicado quando os requisitos se sobrepõem entre eles. Para mitigar esses desafios, as organizações podem realizar estudos de mapeamento ou integração a fim de identificar similaridades e diferenças entre os mecanismos implementados. No entanto, destes estudos podem originar também dificuldades devido à sua complexidade inerente. Na nossa pesquisa, utilizámos a ”Design Science Research Methodology” e realizámos uma Revisão Sistemática da Literatura (RSL) para identificar os principais problemas, bem como desenhar, propor, demonstrar e avaliar uma solução para os mesmos. O principal problema que identificámos durante a RSL foi a complexidade temporal associada ao processo de mapeamento. A nossa investigação tem como objetivo aprimorar o processo de mapeamento através da introdução de uma ferramenta de software projetada para facilitar a fase de comparação do processo de mapeamento. A ferramenta automatiza a comparação de requisitos entre duas normas e gera uma tabela com as sobreposições de requisitos mais prováveis entre elas. A ferramenta foi avaliada ao reunir duas tabelas de mapeamento criadas manualmente, uma derivada da literatura e outra criada por nós, que servem como base de comparação. Comparamos os resultados gerados pela nossa ferramenta com os das tabelas de mapeamento base para avaliar, por meio de várias métricas que propusémos, a precisão dos resultados.

Palavras Chave

Segurança; Ciber-Segurança; Metodologia de Desenho de Ciência; Revisão Sistemática de Literatura; Normas; Inteligência Artificial; Processamento de Linguagem Natural

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

ACM	Association for Computing Machinery
AI	Artificial Intelligence
BC	Business Continuity
BERT	Bidirectional Encoder Representations from Transformers
COBIT	Control Objectives for Information Technologies
CSS	Cascading Style Sheets
DSRM	Design Science Research Method
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HTML	HyperText Markup Language
IDF	Inverse Document Frequency
IR	Information Retrieval
IS	Information Systems
ISMS	Information Security Management System
ISO	International Standards Organization
ITIL	Information Technology Infrastructure Library
JS	Javascript
NLTK	Natural Language Toolkit
PRM	Process Reference Model
RQ	Research Questions

RSC	Risk, Security, Compliance
regex	Regular Expressions
SBERT	Sentence-BERT
SLR	Systematic Literature Review
STS	Short Text Similarity
TF	Term Frequency
WMD	Word Mover's Distances

1

Introduction

Information has an intrinsic value [6] that must be safeguarded from threats. Proper information security measures are therefore necessary to achieve resilience. By implementing such measures, organizations can reduce risks and ensure that their critical information is protected.

Failure to implement adequate information security measures can result in significant consequences and costs [7, 8]. Therefore, organizations must prioritize information security to ensure their long-term viability and success. By doing so, they can mitigate risks and protect their valuable assets, which can include sensitive customer data, confidential business information, and intellectual property [9, 10].

In today's interconnected world, where cyber threats are becoming more sophisticated, robust information security measures are essential for ensuring organizational resilience and growth [11]. However, many organizations have neglected to take advantage of the latest research in cybersecurity to enhance their security posture. Instead, they often prioritize compliance with security requirements [12]. As a result, there is often a disconnect between the latest research produced by academia and the security measures that organizations implement, hindering their ability to effectively protect their critical assets.

The main purpose of this document is to explore strategies for bridging the gap between academia and industry by identifying areas of research that can enhance organizational security, risk management, and resilience. In particular, we aim to examine how organizations can improve their security posture by properly utilizing multiple compliance mechanisms that better align with the evolving threat landscape.

Organisations often need to implement multiple compliance mechanisms into their systems and processes. For instance, European organizations faced the task of integrating the newly introduced General Data Protection Regulation (GDPR) [13] into their existing security operations prior to 2018.

However, compliance with legislation is not the only consideration for organizations seeking to enhance their security posture. They can also choose to implement standards, such as International Standards Organization (ISO) 27001 [2] and Control Objectives for Information Technologies (COBIT)'s DS5 process, Ensure Systems Security, [14], both of which focus on improving IT security. Despite sharing a common goal, these standards have fundamentally different focuses. In particular, COBIT takes a more extensive governance-focused view compared to ISO 27001 [2], which places greater emphasis on operational controls [15]. This phenomenon creates an interest for organisations to implement both standards, to address security from both the operational and strategic points of view.

Organisations often have a vested interest in the implementation of multiple compliance mechanisms. This incentive can be either mandatory, as in the case of complying with new legislation, or discretionary, as in the case of improving organizational processes to achieve strategical goals. This situation presents a complex scenario where organizations need to navigate overlapping requirements and ensure compliance with all relevant compliance mechanisms [16].

Organizations can face challenges when implementing multiple compliance mechanisms without proper planning and coordination. There can be significant overlap of requirements between the mechanisms, which can lead to confusion about what has already been implemented and what has yet to done [15]. Furthermore, adapting an organization's processes to meet the requirements of each mechanism separately can be costly in terms of both time and resources [17]. Furthermore, there can be differences in terminology used to describe the same entity across different international standards. Synonyms are common, especially between standards originating from different organizations or dealing with the same scopes in different contexts.

To address these challenges, it is crucial for organizations to develop a comprehensive strategy that takes into account the unique requirements and potential overlaps of multiple compliance mechanisms. This strategy should include proper planning and coordination to minimize confusion, reduce costs, and ensure compliance with all relevant compliance mechanisms.

This document proposes a strategy to address the challenges of implementing multiple compliance mechanisms. The strategy involves three processes: Homogenization, Mapping, and Integration, which are applied sequentially to create an integrated model that meets the requirements of all the relevant compliance mechanisms. This approach simplifies the implementation process by generating a set of requirements that ensures compliance with all necessary mechanisms

Despite its numerous benefits [15, 17–29], the implementation of this approach of homogenizing, mapping or even integrating standards can be extremely time-consuming [28]. Furthermore, there is a lack of research on how to streamline and speed up the implementation of these processes.

We followed the Design Science Research Method (DSRM) proposed by Hevner et al. [30] in the course of our research. We first started by identifying the existing research problems in the processes of harmonization, mapping and integration of Standards, utilizing the Systematic Literature Review (SLR) methodology for that purpose [31]. Based on the findings from the SLR, we designed, developed and evaluated a tool to assist researchers and practitioners in

the mapping process.

The developed tool takes a document's normative requirements (or ISO Standard document) as input and automatically maps them to another document's normative requirements. The output is a mapping table that simplifies the comparison of multiple compliance mechanisms.

To evaluate the performance of the developed tool, we utilized Information Retrieval (IR) methods and compared its outputs against two existing artifacts, both of which were developed manually. One of these artifacts was developed in the context of this document, while the other was published by Pardo et al. [1].

The main contribution of this document is the developed tool, which aids researchers in the mapping process and demonstrates reasonable precision when used appropriately, particularly for excluding clauses within a document that have no possible mapping to another compliance mechanism.

The structure of this document is as follows: the introduction provides an overview of the document and the field of study. The subsequent sections include Research Background, where the main concepts necessary for understanding our research are presented, and Related Work, which highlights three important papers in the field.Following the Related Work section, we present the Systematic Literature Review, divided into planning, conducting, and reporting subsections. We utilize this information to identify the research problem in the subsequent section.After identifying the research problem, we propose a solution and illustrate its practical application through the development of a software tool, respectively in the Proposal and Demonstration sections. The tool's performance evaluation is discussed in the subsequent section. Finally, the conclusion summarizes the study and its main contributions. Within this section, we address the limitations of our work and outline potential future research directions.

2

Research Background

Contents

In this section, we present the concepts that are essential to understand our research. We split the research background into the three distinct parts that make up our research: compliance, including a brief presentation of what standards are and which ones we used in our research, the three processes, Homogenization, Mapping and Integration, that we studied in depth and text processing methods, with NLP techniques and AI.

2.1 Compliance

A compliance mechanism refers to a model or framework adopted by organizations to adhere to specific requirement or best practices, respectively, aiming to enhance various aspects of their operations. These mechanisms encompass a range of tools, such as legislation, internal policies, and standards. These mechanisms generally follow specific guidelines pertaining to the organization that issued them. For instance, the ISO refers to the requirements of their standards by ”controls”.

These mechanisms are often presented as models, which can make the process of simultaneously assessing them easier. This is especially useful when implementing standards deriving from different fields of research [32]. In this document, we focus on three areas: Business Continuity (BC), Information Security and Risk Management and, in the following subsections, we present the most important ISO Standards related to them and our research.

2.1.1 ISO 22301 - Business Continuity

The areas of BC and Risk Management are closely intertwined. Where Risk Management generally defines how to deal with risks at an individual level, by setting a risk appetite (the amount of risk, an organization is willing to accept in pursuit of value) [33], BC generally takes this a step further and seeks to explore how risk can impact the existence of the organization itself, through the collective set of risks that an organization can expect [34].

The ISO 22301 Standard sets requirements to standardize organizations' BC Management Systems, particularly by requiring business continuity objectives, which must be maintained, measured and continually improved [3]. These objectives can take many forms, from requiring risks to be treated to ensuring that, even in case of systemic threats, the critical systems be maintained to ensure pre-defined minimums. This Standard does not specify how risks should be treated, the reference for risk assessment and treatment is the ISO 31000 Standard [35].

2.1.2 ISO 27001 - Information Security

A key aspect of minimizing risk, and ensuring BC, is to implement proper Information Security measures across an organization [36]. To achieve this goal, many compliance mechanisms have been developed for use within organizations, such as COBIT's DS5 [14] or the GDPR [13].

The main Standard we focus on, for Information Security, is the ISO 2700X family of standards [2, 37]. These Standards follow a two-part system, where the ISO 27001 Standard is the requirements document, which specifies the controls to be implemented and the ISO 27002 Standard is the implementation guidance on how to implement the requirements.

The ISO 27001 [2] Standard specifies that organizations must develop an Information Security Management System (ISMS) for a given scope and policy, as well as to train staff in regards to the newly implemented system. The Standard requires organizations to document any risks or possible incidents and to properly respond to incidents as they appear. The ISMS must not be static, continual improvement must be ensured to keep the system up to date be it through management reviews or internal audits.

2.2 Homogenization, Mapping and Integration of compliance mechanisms

Homogenization involves comparing two different models from both a semantic and structural perspective [38]. It is the first step in the pipeline and is necessary when the models being compared have dissimilar structures or semantics. An ontology can be used to provide a vocabulary and the necessary relationships to make this process easier [39].

Mapping is the process of comparing two different models and presenting the outputs in a mapping table or artifacts [38]. Integration combines the compared models into a single unified model [32]. Table 2.1 provides a list of relevant terms related to homogenization, mapping, and integration.

2.3 Natural Language Processing

NLP is a field that explores how computers can be used to understand and manipulate natural language text or speech [50]. This subsection introduces NLP and Artificial Intelligence (AI), and Table 2.2 summarizes the most relevant terms and their definitions.

Term	Definitions
Harmonization	Activity that seeks to define and to configure the most suitable harmonization strategy for achieving the strategic goals of an organization where two or more models are involved. [40]
Harmonization strategy	A harmonization strategy is a process which is comprised of a set of methods and techniques defined systematically, which allows us to know "what to do", as well as "how to put" two or more models in consonance with each other. The harmonization strategy is the main work product that any harmonization project must obtain to put two or more models in consonance [41]
Mapping	Comparison technique that goes far beyond the simple identification of the differences and similarities between the elements of the models that are compared. [42]
Integration	Action or effect of joining or merging two or more models [43, 44]
Homogenization	Set of steps and tools by which one or more models are treated, to convert the structures of their process elements into homogeneous structures. [45]
Ontology	An Ontology defines what exists for a given field or discipline. It is generally a leveled construct, with categories and subcategories grouping the lower level elements. Ontologies are often used to map between models of the same field, providing keywords and a knowledge basis [46]
Normative requirement	Process or policy to be implemented as part of a standard's requirements.
Control	Normative requirement specific to ISO Standards
Atomic normative requirement	Normative requirement that can no longer be subdivided into parts [47]
Metamodel	A metamodel is a model that consists of statements about models [48]
Process Reference Model (PRM)	A process reference model helps to define a set of processes which support objectives of a domain, and has two components: domain and scope, and purpose and process outcomes [49]
Coverage	Percentage of the atomic controls covered for a given construct
Directionality	The Directionality of a mapping determines if the mapping is computed from all standards to all mapped standards
Scalability	Ease of extension

Table 2.1: Terms related to modelling and standards and their definitions

An NLP pipeline typically begins by splitting a document into sentences and then into tokens. The resulting data is used to build an inverted index, where the keys are the vocabulary, and the nodes point to documents from the corpus. Tokens can be further processed using lemmatization or stemming to remove derivational suffixes [51]. N-grams can also be parsed from the text and added to the inverted index.

The inverted index provides statistical data on which documents are more relevant for a given query, using metrics such as Term Frequency (TF) and Inverse Document Frequency (IDF). TF matches terms present in the query and documents in the corpus, while IDF negatively ranks common terms across the corpus. These two metrics are multiplied to generate the TF-IDF score, which is a classic method for ranking documents, that has proven its efficacy over and over again for over 40 years [52, 53]. TF-IDF values for two different documents can be put into a pair of vectors, and the cosine similarity between them can be calculated to obtain similarity

scores.

2.3.1 Sentence Similarity through AI

Deep learning has applications across various fields, such as reverse image search and speech recognition [54]. One particular application that is relevant to our research is sentence similarity and text transformers, which use text encoders.

Text transformers were introduced in 2017 as a method to speed up older methods of Deep Learning, such as Recurrent Neural Networks, by allowing them to take all the data in at once, enabling easier parallelization [55].

These models are based on a pair of Encoder and Decoder. The encoder converts the inputted text into a vector of floats, representing the ”value” of each word within a given context. Generally, this vector is referred to as ”embeddings” or ”encodings”.

Text encoders can be used to evaluate text similarity, by using their embeddings [56]. Decoders, on the other hand, take these embeddings and additional text input to generate more text, which is useful in applications such as chat bots like the Generative Pre-trained Transformer (GPT) family.

Although transformers, like Bidirectional Encoder Representations from Transformers (BERT) brought significant speed improvements, it is a generalized encoder and, as such, is slow, even on modern hardware, to finish pair regression tasks like sentence similarity [57]. Reimeirs et al. addressed this issue in a 2019 study that specifically addressed sentence similarity, creating Sentence-BERT (SBERT), achieving measured speedups of up to 780 times [57].

Term	Description
Information Retrieval IR	The process of analyzing text and identifying mentions of semantically defined entities and relationships within it [58]
Document	Unstructured text input
Lemmatization	Lemmatization removes inflectional and attempts to return the dictionary form of a given token [51].
Stemming	Stemming is a procedure to reduce all words with the same stem to a common form [51]
Inverted index	Mapping of each token to which documents contain it [59]
Token	Word within a vocabulary
Vocabulary	Set of unique tokens
Corpus	Set of documents
n-gram	Expression containing n tokens
TF	Term Frequency
IDF	Inverse Document Frequency

Table 2.2: Common NLP terms and their definitions

3

Related Work

Contents

In this Section, we discuss the related work already published in this field, how our document differs from the literature and why we believe that our research brings valuable contributions to the fields of organizational security and resilience.

Our main goal with this report is to improve organizational resilience by making it easier to implement multiple compliance mechanisms within a given organization. We began by analyzing the existing literature over the harmonization, mapping and integration of ISO/IEC Standards.

3.1 Work related to compliance

In 2020, researchers carried out a Multivocal Literature Review, in which the authors present their findings over the mapping of security standards across five sources (two of which were non-scientific) [60].

The paper concludes that, despite the existence of a large number of mappings, the current mapping methodologies are limited and should be further researched, especially due to the importance of present-day Security standards. To solve some of the issues on the field, the authors suggest the implementation of NLP techniques to assist in the mapping of standards, but to the best of our knowledge, published no further work

Besides being older than our review, [60] differs from ours in three ways:

- We did not consider non-academic sources and instead considered the Scopus database, a reliable source [31] that is not included in [60].
- We took into consideration not only mappings, but also integrations of standards
- We considered all ISO or IEC standards related to risk, business continuity and security, not just security

The study present in [61] is a comparative analysis between only two integrations. The authors reached the conclusion that integrating ISO/IEC 27001 [4] with COBIT [14] and Information Technology Infrastructure Library (ITIL) [62] reach similar benefits. Both integrations increase credibility of Information Security, but where COBIT raises credibility on the Governance side, ITIL does so on the IT management's side.

Due to the much narrower scope of [61] when compared to ours, we conclude that the studies share similar goals, but take different approaches. While their paper brought forward

detailed information over two specific integrations, ours focuses on raising the State of the Art across all the techniques used in the industry and academia to map and integrate ISO or IEC standards related to risk, business continuity and security.

Moving on to a Literature Review carried out by Pardo et al. [63] which covers the same subject as the present document. We assess that this paper is extremely thorough in its research methodology.

The paper notes that the idea of studying mappings between these kinds of mechanisms is not new but has been gaining traction. The authors of the study also note that future work is required within the field to improve the process of harmonizing between multiple models and propose that a framework would be useful in this regard.

Two of these studies have somewhat limited scopes, be it in what kinds of processes they opt to include [60], or in the number of artifacts analyzed [61]. The third paper we considered is now fairly dated, as it was published over a decade ago by the time of writing [63]. As such, we opted to conduct our own Systematic Literature Review (SLR) on this topic to identify any new work developed in the field.

3.2 Work related to Sentence Similarity

In this section, we discuss two papers related to natural language processing in our studies. The first paper, published more recently, focuses on an older technology but contributes to the development of a systematic literature review (SLR) to advance the State of the Art on Short Text Similarity (STS) [64]. The second paper, on the other hand, explores the use of neural networks, a modern technology, to achieve comparable or superior results to classical NLP methods [57].

The paper by Prakoso et al. (2021) presents an SLR that leverages NLP techniques to improve the State of the Art on STS [64]. The authors of that study propose three distinct techniques, which can be combined as a hybrid approach, for measuring text similarity:

- **String** - Utilizes only the string's contents to generate a score.
- **Knowledge** - Utilizes string's contents and a semantic knowledge base to generate a score. This allows the algorithm to use, for example, synonyms of words and a given weight to improve the similarity rankings.
- **Corpus** - Utilizes the string's contents and a large corpus to provide semantic insight [65].

Where Prakso et al. picked up on the classic methods to apply STS [64], Reimeirs et al. opted to evaluate, and develop, new methods of carrying out these processes, namely through the use of neural networks [57].

The authors of the second study propose that the use of STS with the networks ”BERT” [56] and RoBERTa [66] is too slow for use in practice. As such, they improved on the processes through the application of siamese neural networks to BERT and RoBERTa, achieving major speedups at negligible hits in the accuracy of the network.

After careful review of these two main studies, we have found that their scope is large enough to not warrant an additional Literature Review on STS, especially considering that both of them are recent studies, published within the last 4 years.

4

Research Methodology

Contents

4.1 Design Science Research Methodology

We applied DSRM to the field of Information Systems (IS) following the guidelines presented by Hevner [30]. To identify the Research Problem, we opted to carry out the Systematic Literature Review methodology described in the following Subsection. We chose this process to raise the State of the Art on the Homogenization, Mapping and Integration of ISO Standards due to the rigor of the process, which results in credible and trustworthy research artifacts.

The main problem we identified was that the process of mapping between compliance mechanisms, such as international standards, legislation or internal policy, was time-consuming and cumbersome for researchers. To address this challenge, we proposed the development of a software tool to automate certain repetitive aspects of the process. We followed the DSRM process by designing and creating the tool, and generating performance metrics based on literature artifacts. The tool, being the research artifact, was continually improved throughout three iterations.

Finally, we communicated our findings for peer review, completing the DSRM process. Figure 4.1 summarizes the guidelines we followed in our research.

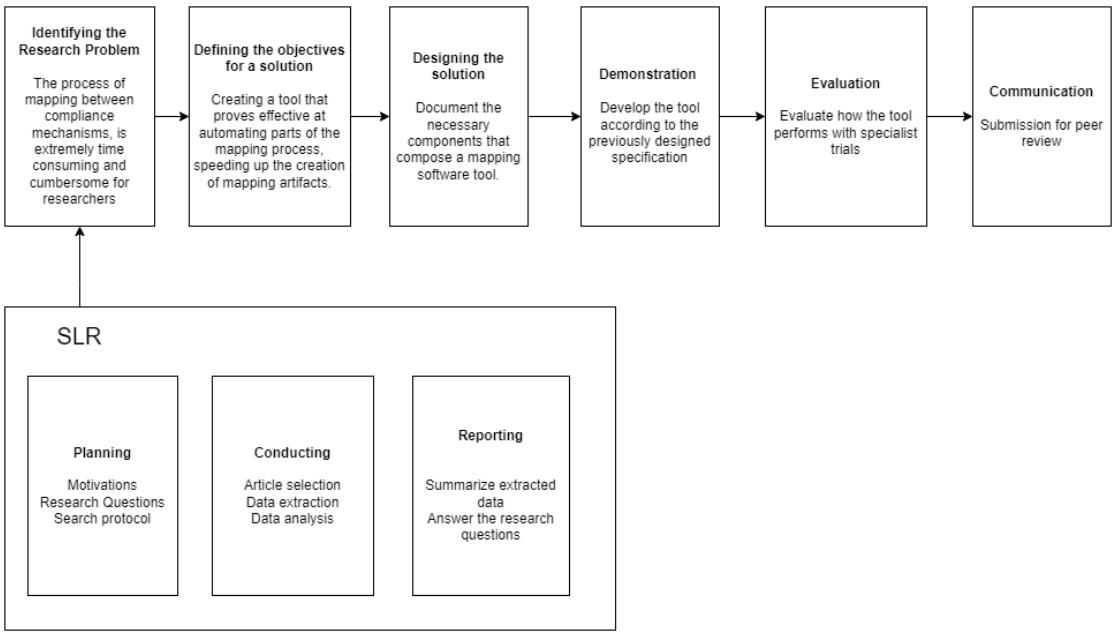

Figure 4.1: Summary of the DSRM guidelines applied to our study

4.2 Systematic Literature Review

A SLR is a research methodology that aims to produce a fair evaluation of a research topic through trustworthy, rigorous, and auditable means [31, 67]. We carried out a three-stage process composed of the steps proposed by Kitchenham [31].

The first stage, planning, includes the motivations for our work, the Research Questions we aimed to respond to and the reasoning for them, and the search protocol we designed and followed to find the articles we would pick and read through.

Secondly, we carried out the conducting phase, which consists in the execution of the search strategy. In this stage, we note how many articles were found and how many of those were included or excluded.

The last stage is the reporting phase, in which we respond to the questions we drafted in the conducting stage.

5

Systematic Literature Review

Contents

5.1 Planning

In the Planning stage of this SLR, we describe our methodology, including our motivations that led to the beginning of the paper, the Research Questions (R.Q.) that we had aimed to respond to and the search protocol we drafted to find and review the articles for our paper.

5.1.1 Motivation

To improve internal processes or even assist in reaching compliance with existing regulation [68], the route organizations often pick is to seek certification in some international standards such as ISO/IEC. However, the literature shows that implementing multiple of these unintegrated standards can lead to reduced gains in performance with each additional implementation [69].

At the same time, some research has focused on analyzing the gaps between standards so as to reduce the amount of work needed to implement them [60, 61]. However, the research always focuses on either mappings or integrations, but never both. As such, we decided to review the State of the Art on both of these processes with a single SLR.

5.1.2 Research Questions

In the following paragraphs, we present the Research Questions we respond to during our literature review, with the first three focusing on the metadetails of mappings and integrations, while the latter two instead focus on the artifacts generated from their respective processes.

We aimed to discover if the literature presented the mapping and integration of standards as a useful process, both in the industry and academia. To achieve this goal we drafted Research Questions (RQ) 1, in which we aimed to identify the benefits deriving from the mapping and integration of standards.

With R.Q. 2, we meant to discover any challenges during the mapping or integration of standards. We reasoned that finding challenges in these processes might indicate gaps in the literature or existing problems that could be addressed with further study.

The goals for the drafting of R.Qs. 3-5 are aligned. With this set of questions we aimed to discover any patterns in what kinds of artifact are used for mapping/integrating which standards and what kind of mapping the research produced. Moreover, we also seeked to discover if any mapping/integration utilized software to assist in the process.

R.Q. 1 - What benefits exist from mapping or integrating standards?

R.Q. 2 - What challenges derive from mapping or integrating standards?

R.Q. 3 - What kinds of mappings and integrations exist for standards?

R.Q. 4 - Which artifacts have been proposed for mapping or integrating standards...

R.Q. 4.1 - ... at a complexity level higher than the standard's?

R.Q. 4.2 - ... at the standard's level of complexity?

R.Q. 4.3 - ... at a complexity level lower than the standard's?

R.Q. 4.4 - ... utilizing software?

R.Q. 5 - What standards have been mapped and using which artifact?

5.1.3 Search Protocol

With the goal of finding all the relevant articles related to the mapping and integration of ISO/IEC standards and focusing on Risk, Security, Compliance (RSC), we carried out our search in Scopus' and Association for Computing Machinery (ACM) digital databases due to their credibility [31].

The search string used to find the articles is: **(mapping OR integration OR integrating) AND (risk OR security OR "business continuity") AND (ISO or IEC)**. We built the string by selecting first the processes we aimed to research (Mapping and Integration), the scope of research (RSC) and the types of standards we aimed to include (ISO/IEC).

Initially we had considered widening the scope of the SLR by removing the RSC limitation from the search string. However, this attempt resulted in the number of articles emanating from the search to increase substantially. Due to the limited number of resources on our research group, we opted to maintain the limitation to our group's field of expertise.

5.1.3.A Inclusion and Exclusion Criteria

To make a decision on whether to include or otherwise exclude a study from those found by searching the database with the search string, we devised a set of exclusion and inclusion criteria to guide the study selection process [31]. These criteria were defined before beginning the processing of articles in an effort to minimize researcher bias. When any inclusion criteria is not met, the study is excluded (as would be the case, for example, with all studies in a language other than English or Portuguese).

Inclusion Criteria

Source - Source material is a book, journal or conference proceeding

Language - Written in English or Portuguese

Type - Is a primary study or reporting on a primary study

Field - Is related to business continuity, risk or security

ISO/IEC Standard - Includes the mapping or integration of at least one ISO or IEC standard.

Exclusion Criteria

Duplicate - Article is a duplicate of another

Accessibility - Source not available for the full text

5.2 Conducting

During this stage, we executed the search strategy defined in the planning phase to identify a grand total of 718 papers as of the export date of the 29th of November of 2022. Of these 718 papers, 32 were identified as duplicates and were removed from the selection, leaving us with 686 papers. These would be classified as ”Included”, ”Excluded” or ”Maybe” in the following steps. A summary of the exclusion process can be seen in Figure 5.1.

5.2.1 Search and Selection Proceedings

After reading the titles and abstracts, we excluded 606 papers. We have found that the inclusion of the term ”integrating” in the search string resulted in a large number of search results related to the integration of a standard in organisations, which is clearly off-scope based on our exclusion criteria, namely, the fact that they are not mappings or integrations of two or more standards, but rather the implementation of a single one.

In the following step, we decided to read through the introductions of the 55 ”maybes” to make a decision on their inclusion or exclusion. This lead to the exclusion of 49 papers and inclusion of 6.

We read all 36 included papers' full text and excluded 11 more, accepting the set of the remaining 25 papers as final.

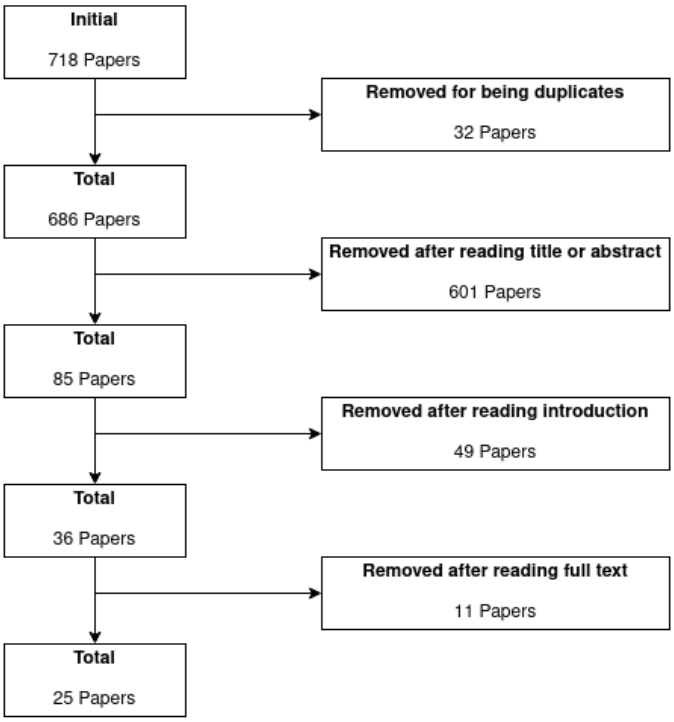

Figure 5.1: Summary of the selection process

5.3 Reporting

At this stage we present and discuss the answers to the previously proposed research questions with information based on the reviewed literature.

5.3.1 R.Q. 1 What benefits exist from mapping or integrating standards?

The literature states that there are benefits deriving from standards' mapping and integration [15, 17–29]. A large part of these benefits are based on improved access and quality of information [25, 27], allowing for more efficient communication of information. A summary of these benefits can be found in Table 5.1.

Mappings can help bridge the gap between different areas of expertise [20, 25, 29]. For example, the literature proposes that gap between Governance and Information Technology can be bridged through a mapping between ISO 27002 and COBIT's DS 5 - Ensure Systems Security help [25, 29], with the ISO standard presenting the "how" (the technical aspects at a lower level) and COBIT the "why/what" (the higher level, governance side aspects). This is

Table 5.1: Summary of the most relevant benefits across the literature

	Better Auditing and Compliance	Improved Cooperation	Improved Risk Management	Higher Efficiency	Easier Standard Adoption
[23]	X	X	X	X	X
[25]	X	X			
[29]	X	X			
[28]	X		X		
[19]	X				X
[20]	X				X
[27]		X			
[32]		X			
[24]			X		
[15]			X	X	
[17]				X	
[22]					X
[26]					X
[21]					X

especially helpful since it improves communication, reusability and organization of knowledge [27].

In addition to a more efficient flow of information, law, policy and standards' compliance is also shown to be significantly improved with the mapping and integration of standards [19, 20, 23, 25, 28, 29]. There are two main improvements that lead to this benefit:

- **Easier implementation -** By having all the information of what controls need to be implemented in one single artifact, enterprises have an easier time determining what controls have already been implemented and maintained [20].

- **Better auditing -** Auditing is another essential aspect of compliance [70] that can be improved by the mapping and integration of standards. The literature shows us that it is easier for auditors to validate compliance [23, 25, 28], leading to better law compliance [20] and reduced costs from more efficient use of internal auditors' time [22].

We have found evidence that indicates that Risk Management is another area that benefits from the mapping and integration of standards [24], often through mitigation of risk, causing a reduction in the number of incidents across the same timeframe [28].

The literature also provides evidence that the business itself can benefit from the mapping and integration of standards. Certification is easier to attain [21, 23], primarily when one of the mapped standards has already been implemented [26].

Moreover, organizations can observed improved public image through the implemented cer-

tifications [22] and transparency [17], making it easier to attract new customers or improve the loyalty of existing ones [23].

5.3.2 R.Q. 2 What challenges derive from mapping or integrating standards?

The literature presents a sizable lack of information regarding the challenges present in the process of mapping or integrating standards. We believe that is due to a lack of systematic reporting, leading to the articles being entirely focused on presenting the results of their findings and not so much on the process used.

The few challenges found are mainly related to the artifact used to map the standards, making generalizations hard to derive from the literature. However, we can say that mapping standards is often a very time consuming process. Most mapping methods involve multiple review sessions where researchers meet, discuss ideas, share knowledge improve upon the proceedings of the previous meetings [71].

Taking a look at a specific artifact, ontologies, we can derive that it is often impossible to map onto them bi-directionally with relative term accuracy. Comprehensive ontologies include terms from multiple standards and as such, will not be able to be fully mapped back onto a single standard. Thus, one must make a choice between bi-directionality of the mapping and coverage of terms on the ontology, which can be a challenge on its own [27].

One challenge we discovered that is not directly presented in the articles but can we can infer from the literature how resource intensive the mapping and integration processes can be, using up a significant portion of a group of researcher's time [28].

5.3.3 R.Q. 3 What kinds of mappings/integrations exist for standards?

The literature does not present different kinds of integrations. However, every integration requires a mapping to be done beforehand. Thus, we are only considering mappings to answer this question as the answers given also applies to the mapping stage of integrations. To assist in understanding and responding to this research question, we bring to light two distinct concepts: abstraction level and directionality:

Abstraction level - mappings can engulf concepts at different abstraction levels and as such, the relationships between elements can change. It is possible to map a single control to many [25], many to one [27], as is usually the case with ontologies [27, 72]. It is also possible to

map elements other than controls, including categories or processes, provided that all of their lower level controls are mapped to [71].

Directionality - Unidirectional mappings present a one-way map: corresponding the terms of one standard to terms of another, but not the opposite. This property leads to more straightforward mappings by reducing artifact sizes, but cuts some use cases. Bi-Directional mappings encompass all use cases by ensuring that either standard is both source and destination and allowing elements to be mapped starting from any of its mapped standards. In some cases, researchers claim for it to be trivial to extend a Unidirectional mapping by retracing its steps [25].

5.3.4 R.Q. 4 Which artifacts are used/have been proposed for mapping or integrating ISO/IEC standards?

We decided to group each of the artifacts proposed by the literature into one of four groups, based on the complexity of the artifact: Group 1 - Construct (frameworks, ontologies, metamodels) Group 2 - Model Group 3 - Method/Algorithm Group 4 - Software (automated)

In Group 1, we discovered two frameworks: SABSA [20] and HFramework [26] and two unnamed security ontologies [27, 72]. These generally aim to generalize and map standards onto them. The ontologies have great coverage of the standards they aim to span.

In Group 2, we found some models created by researchers to map standards [18, 32], in specific, Process Reference Models (PRMs) been used to map atomic requirements in a stricter way [21].

Group 3 presents methods with less formality than the previous ones, replacing the well defined (meta)models with an algorithm or list of methods. Some follow stricter guidelines like [19, 71], others a more lenient approach, using a sequential list of strategies [24].

Group 4 would present entries related to mappings and integrations deriving from software, but we have not found any in the literature.

5.3.5 R.Q. 5 What standards have been mapped and using which artifact?

We identified that ISO 27000 comprised most of the mappings and integrations using ISO standards, amounting to 54% of the total identified mappings and 33% of integrations.

Below, in Tables 5.2 and 5.3 we present a table containing the most commonly mapped (Table 5.2) and integrated (Table 5.3) standards and which group of artifacts they were processed with. The groups were distributed in the same manner and logic as in R.Q. 4. In the columns,

we list the aforementioned groups of artifacts and in the rows, the most common standards. It is worth noting that we grouped the entire family of ISO 27000 standards into a single row (ISO 2700X), including ISO 27001, 27002 and 17799 [2, 4, 37].

Table 5.2: Most relevant mapping data

Mapping	Construct	Model	Method	Software	**Total**
ISO 2700X	3	4	5	0	12
ISO 20000	0	2	1	0	3
ISO 15504	0	0	1	0	1
ISO 9001	0	1	0	0	1
COBIT	0	3	2	0	5
Total	3	10	9	0	22

Table 5.3: Most relevant integration data

Integration	Construct	Model	Method	Software	**Total**
ISO 2700X	1	1	0	0	2
ISO 20000	1	1	0	0	2
ISO 15504	0	1	0	0	1
ISO 9001	0	1	0	0	1
COBIT	0	0	0	0	0
Total	2	4	0	0	6

From the data we can gather that there is a significant interest in the ISO 2700X family of standards, spanning over half of the mappings and a third of all integrations. COBIT also appeared in a large amount of mappings despite not being specifically included in the search.

We can also conclude that around half the authors opted for formal mapping and integration methods, using constructs and models, while the other half preferred less formal methods. It is also worth noting that no mappings or integrations utilized software to automate processes.

6

Research Problem

Our SLR has shown that the mapping and integration of ISO/IEC standards are very beneficial processes for organizational development. They allow organizations to have an easier time adopting new standards while improving compliance mechanisms and reducing implementation costs. However, our review has also identified significant challenges associated with the current methods of generating mapping artifacts.

One major challenge is the time-intensive nature of the mapping process. This can pose difficulties in resource-constrained situations, as it requires a group of researchers to engage in iterative meetings and discussions to determine the corresponding parts of each standard. The outcomes of these discussions are then analyzed to create a mapping artifact, typically in the form of a mapping table. Additionally, the intermediate information generated during the mapping process, including the thought process and the notes taken, is not traceable back to the artifact and often goes unreported. Consequently, if any changes are necessary, the entire time-consuming process must be repeated to modify the artifact.

Changes to standards' mappings can originate from a few different sources. Below we explore some of the possible needs that can lead to the redoing of parts of a mapping/integration.

- **Revision -** Standards are often revised and updated. Sometimes these revisions can cause major changes to the controls or structure
- **Directionality -** It is possible to create a directional mapping, but leave open the possibility of making it bi-directional in the future
- **Scalability -** Most of the mappings and integrations found in the literature included only two standards, but it is possible to expand them to 3 or more.

Seeing as any two standards can be mapped, provided that there are some common points between them, one can observe the sheer magnitude of combinations that can be formed involving ISO standards. This means that it is not feasible to map every combination of standards manually, due to how time-consuming it would be to create and maintain the artifacts.

Despite the benefits of the current methods of generating mapping artifacts, our SLR has found that there is not a standardized way to map or integrate standards, and there are numerous different methods of achieving the same mapping (and possible integration) between two or more standards [71]. We believe it would be useful for organizations to know what the best approach is on how to adapt their systems to comply with a new standard or certification.

Our SLR has found that the majority of mappings take an iterative design approach. With meeting after meeting, the mappers share ideas and discuss each control's relation to another.

This can be repeated between different researcher groups to achieve less bias or simply have everyone discuss in a ”round-table” approach. However, we do not see this as a scalable process, for it has to be repeated every time a new standard is mapped, resulting in a large time usage.

Some papers utilize computer assistance [25, 29], but only as a database provider. In addition, these studies predate all the others by at least five years, making them reasonably outdated as of this paper's writing. Moreover, the authors are reporting on a mapping from ISO 17799 to COBIT, the former of which is not only outdated, but also “defunct”, having been replaced by the newer ISO 27002, that is the set of guidelines of the certifiable ISO 27001 for Information Security [4].

To summarize, we argue that current methods of generating mapping artifacts are not only cumbersome to execute, but also ineffective as they output a single artifact without the reasoning behind it. This makes future changes to the mapping more complex, since the original thought process behind each control's mapping is not reported on.

The literature also does not present any software-based harmonization, mapping or integration assistance (or automation) tool to assist the researcher by automating some of the time consuming parts of these processes. Since these are often repetitive and time consuming processes, the authors considered that the automation of certain processes could be plausible.

7

Research Proposal

Contents

7.1 Introduction

In this section, we describe our proposed solution that aims to tackle the primary research problem outlined in Section 6. Our goal is to provide an effective and efficient approach that minimizes the substantial time required for the generation of a unified model, as presented in Figure 7.1.

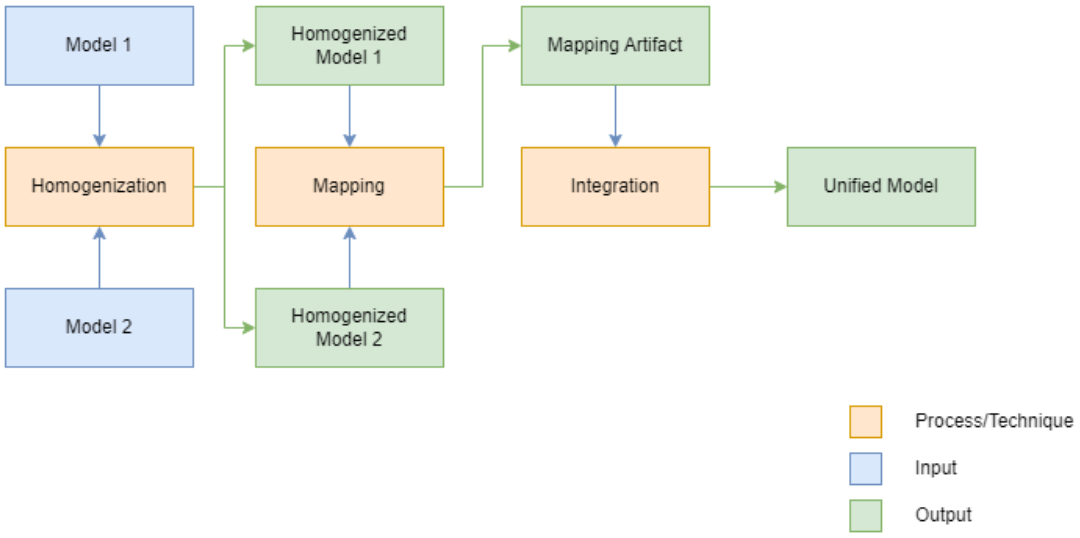

Figure 7.1: Steps to create a unified model [1]

To accomplish this, we propose the development of a software-based tool that automates various aspects of the Homogenization and Mapping processes.

The software-based tool comprises two modules, with one of them focusing on the Homogenization process, while the other leverages the output from the former to facilitate the Mapping process. Naturally, these modules are referred to as ”Homogenizer” and ”Mapper” for the remainder of the present document.

7.2 Homogenizer

The main purpose of this module is to create a reusable homogenization artifact from a compliance mechanism’s document, which can then be used by the Mapper. The artifact is built by listing every atomic control within each section (and subsection) of the document’s structure.

Our research focuses on ISO Standards, so the Homogenizer includes a dedicated mech-

anism for homogenizing these standards. Additionally, we have developed a more general method that can homogenize any type of compliance mechanism.

7.2.1 Homogenizing ISO Standards

The primary advantage, as identified by the authors, in segregating the homogenization process of ISO standards, stems from the ability to capitalize on the internal structure of these standards. A comprehensive analysis of these standards reveals that a majority of ISO controls are already designed as atomic controls, consisting of single statements or easily extractable from well-defined and clearly delineated enumerations.

We propose that the homogenizing system should take the ISO standard's PDF file as input and output a mapping of its clauses to a list of atomic controls. To better illustrate the extraction process of atomic controls, be it manual or automated, a detailed demonstration is provided in Table 7.1.

The extracted atomic controls should be easy to review and edit in an efficient manner, preferably with built-in functionality within the tool itself or with an alternative the average researcher would already be used to working with.

By implementing automation in the extraction of atomic controls, it is possible to significantly enhance the efficiency of the homogenization process. The primary reason for this improvement is that reviewers would be required to spend considerably less time scrutinizing the table generated by the Homogenizer, as opposed to the substantial amount of time they would need if they were tasked with creating such a table manually. This optimization of the process allows for a more effective allocation of resources and an overall improved experience for the reviewers involved.

Original control	Atomic controls
The organization shall: a) identify the risks of disruption to the organization's prioritized activities and to their required resources; b) analyse and evaluate the identified risks; c) determine which risks require treatment	The organization shall identify the risks of disruption to the organization's prioritized activities and to their required resources;
	The organization shall evaluate the identified risks
	The organization shall analyse the identified risks
	The organization shall determine which risks require treatment

Table 7.1: Example of the conversion of a control into its composing atomic controls (ISO 22301) [3]

7.2.2 Homogenizing other types of compliance mechanisms

We argue that creating a homogenization system that only applies to a single class of compliance mechanisms would be a design flaw. Therefore, we propose an alternative module that accepts any list of tuples matching section numbers to atomic normative requirements.

In practice, this Homogenizer should produce the same output as the one proposed in Section 7.2.1. However, the researcher would have to do additional work in identifying each atomic control and storing the data into a spreadsheet. We do not see this as a major constraint because this process is used even for mappings made manually [47].

7.3 Mapper

We propose that the main contribution to the field would derive from improving the mapping process. As highlighted in Section 6, it is the numerous comparisons between the atomic controls of two different compliance mechanisms that takes the bulk of the time spent in creating a mapping.

The mapper would therefore step in by automating the most repetitive part of the mapping process, the comparisons between atomic controls. This module would evaluate each and every pair of atomic normative requirements and output a similarity score between their contents.

In addition to speeding up the mapping process by generating a preliminary mapping table, such a tool could provide some error checking properties when used to compare against pre-existing mappings.

We propose to split this module into two parts: one using classical NLP methods while the other utilizes experimental pre-trained neural networks to compare sentences. The mapper can be easily parametrized with weights to value each method independently.

7.3.1 Mapping with classic NLP techniques

Classical NLP techniques, such as cosine similarity through the use of TF-IDF, act as a baseline for the given tool, as they are now a ”tried and true” method that was proposed in 1988 [53] and has been in use until the present day [73, 74].

When applied to our research problems, we propose a NLP pipeline to assist in the comparison process, which we summarized in Figure 7.2.

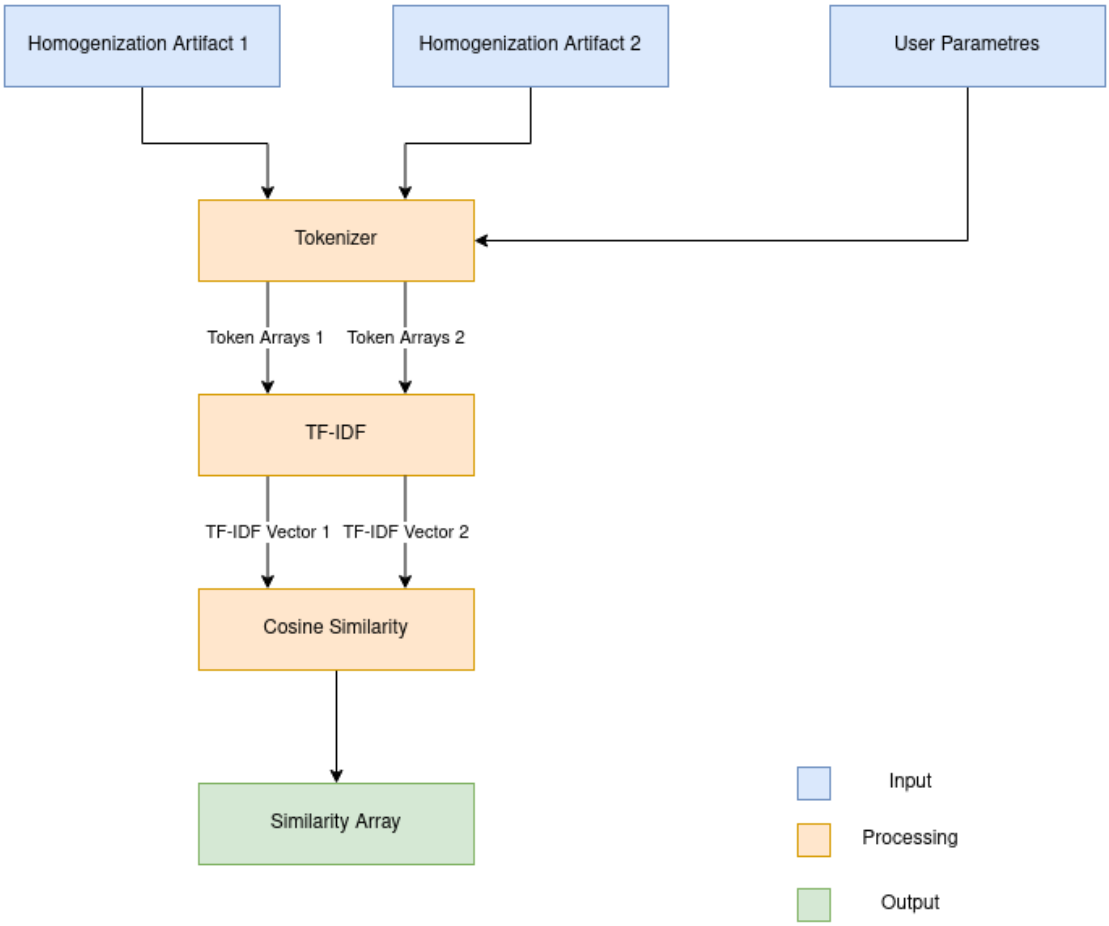

Figure 7.2: Mapping with an NLP pipeline

In order to perform the mapping process, we must determine which controls are the most similar. To achieve this goal, we propose to calculate the cosine similarity of the TF-IDF values between the two artifacts.

There are multiple alternatives to achieve the ”raw” TF calculation presented. We opted to use the ”log-TF” variation, given by the formula `log(1 + TF)`, where TF is the ”raw” TF value. The basis for this choice, is to lower the risk of bias from controls containing a lot of term repetition.

To calculate the value for the IDF, we applied the smoothed out formula for the classic IDF calculation. In this scenario, N is the total number of documents within the corpus and #t is the number of documents in which the term t appears.

$$log(\left(\frac{N}{\#t+1}\right)+1$$

At last, TF-IDF is the simple multiplication of TF by IDF for a given term, in a given document from a given Artifact. This gives us the relevance of a token in relation to its context. By calculating these values to all of the controls in either of the two Preprocessor Artifacts, we gain a base we can use to compare them.

To compute the cosine similarity(csim), we first suggest utilizing a N x M zero matrix, with N equaling the number of controls for the Preprocessor Artifact 1 and M equaling the same for Preprocessor Artifact 2.

$$\begin{bmatrix} csim(n_1,m_1) & \dots & csim(n_1,m_M) \\ \vdots & \ddots & \vdots \\ csim(n_N,m_1) & \dots & csim(n_N,m_M) \end{bmatrix}$$

The result is a similarity matrix, with higher values for a given cell meaning to a higher likelihood of a mapping between the denoted controls.

7.3.2 Mapping with neural networks

We have opted propose the use of the most recent developments in Machine Learning, as a method to improve the results deriving from the classical, NLP, methodology.

The chosen neural networks would have to be pre-trained and their goal would be to perform the ”processing” parts of the mapping process, namely, the pair-wise comparisons between atomic controls.

The overall process would be similar to the one proposed with NLP, but instead of performing the cosine similarity comparisons with arrays, the output array would instead be filled out, cell by cell, utilizing the output of the neural network with the two controls as inputs.

7.3.3 Merging the results

Our mapping proposal aims to produce two distinct metrics out of the same input: one derived from the cosine similarity between TF-IDFs and another from the similarity produced by applying a comparison algorithm provided by a neural network.

The two outputs produce results within the same counter domain: $[0,1] \in \mathbb{R}$, where 0 and 1 denote the minimum and maximum similarities, respectively. However, there is no guarantee

that the results will follow a similar distribution, for the distribution of results of the model's output depends entirely on the architecture of its model and the data it was trained on.

We propose that the NLP and AI similarity values should be normalized to facilitate their comparison. In this case, we propose utilizing the z-score normalization, due to the simple, but effective nature of the process [75].

After their normalization, we can finally merge the two resulting matrices. To do this, we propose grouping the clauses by section and performing any further comparisons on the resulting matrices. This allows us to map the matrices into two objects of the form: $(clause1, clause2) -> similarity_between(clause1, clause2)$, one for NLP and another for AI generated values.

These objects' values are simple vectors we can manage. Thus, we propose the selection of a pair of weights, W1 and W2, such that $W1 + W2 = 1$. These weights can be applied to the similarity vectors V1 and V2 to fulfill a weighted average following the formula: $V3 = W1 * V1 + W2 * V2$. This way, we can configure the weights to value more, or less, each part of the inputs, according to the researcher's preference.

8

Demonstration

Contents

8.1 Introduction

Introducing our developed system, which can be divided into three smaller systems: the front-end, the webserver, and the database provider. A detailed overview of the interactions between these systems is included in this section, with a simplified summary available in Figure 8.1.

The front-end of our system is developed using HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and Javascript (JS), providing a user-friendly interface to receive inputs. The user can make various requests through this front-end, which are handled by the web server.

The web server acts as an interface between the front-end, the database provider, the templating engine, and the three developed modules: (ISO) Extractor, Preprocessor (Homogenizer), and Mapper. Its primary responsibility is receiving and responding to requests, processing the data using the receiving module, and returning the processed data.

The MySQL Server acts as the database provider, receiving and responding to data queries sent by the web server. The database provider ensures data persistence throughout Main Server block restarts, enabling seamless data management.

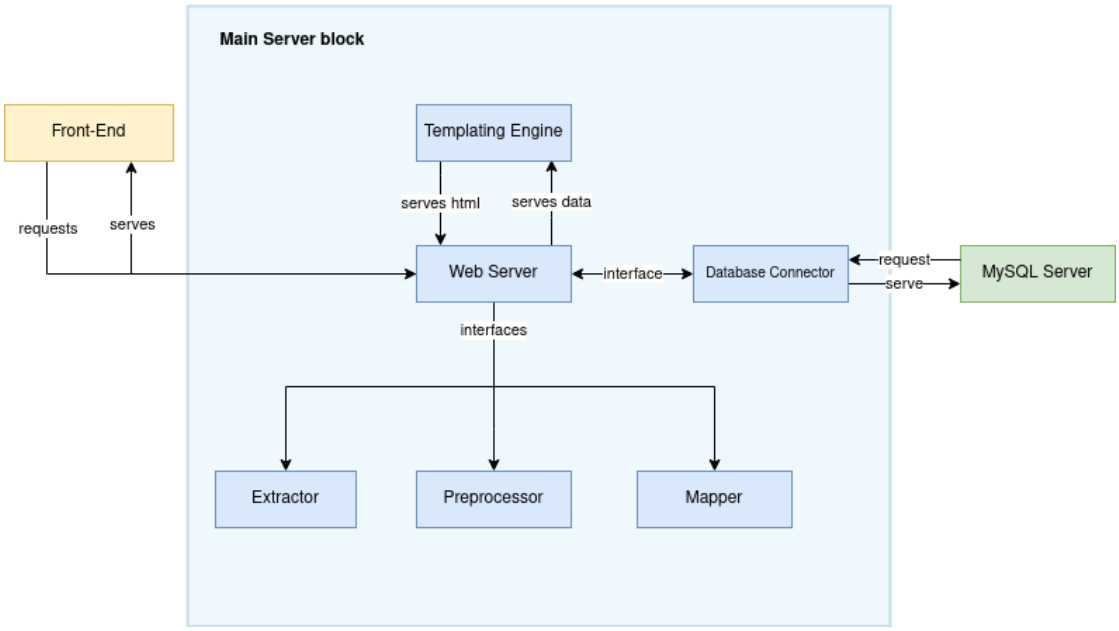

Figure 8.1: System Architecture

8.2 Front-end

The Front-End is built through rendered HTML files, styled using CSS and made interactive through JS. The Front-End navigation is divided into two pages, one for interacting with the Extractor and Preprocessor and another for interacting with the Mapper.

8.2.1 Extractor & Preprocessor

8.2.1.A Home Page

The Extractor & Preprocessor home page, shown in Figure 8.2 gives the user a dashboard that displays available documents, allows the upload of new ones and permits the creation and edit of each document's Preprocessor artifact.

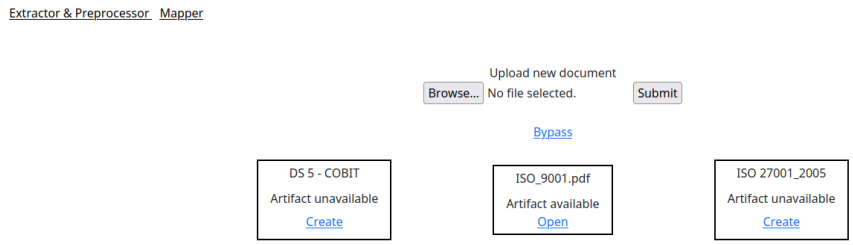

Figure 8.2: Overview of the preprocessor

8.2.1.B Preprocessor Artifact Editing

The Extractor & Preprocessor Artifact Editing page allows the User to review, edit and delete the existing controls for the artifact selected from the Home Page. An example of the editing process can be seen in Figure 8.3

8.3 Main Server Block

The Main Server Block manages all the information that runs through the system. It acts as an interface between the User's requests and the information needed to satisfy them, stored in the MySQL Database. The Main Server Block is composed of a Web Server, the central block, which interfaces with the Templating Engine, The Database Connector and the three data processing modules: The Extractor, Preprocessor and Mapper.

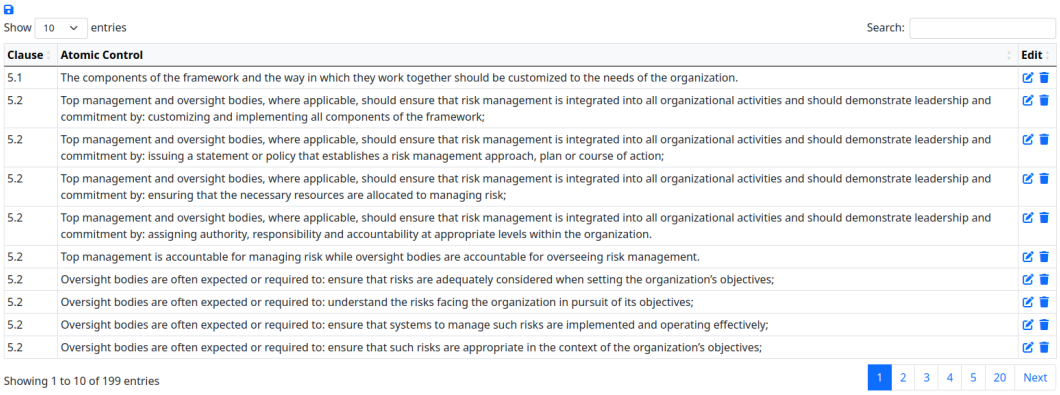

Figure 8.3: Editing an existing Preprocessor Artifact

To develop the Main Server Block, we utilized multiple external modules, some from Python's standard library, others installed through Python's package manager. A summary containing all of our imports per each of our developed modules can be found in Table 8.1

	Extractor	Preprocessor	Mapper	Database Connector	Web Server	Templating Engine	Libraries
collections			x		x		
dotenv [76]				x			
functools					x		
flask [77]					x		
ginja2 [78]						x	
math			x				
mysql [79]				x			
nltk [80]		x	x				
numpy [81]			x		x		
os	x	x			x		
pandas [82]			x		x		
pathlib	x						
pdfplumber [83]	x						
pickle	x	x					
re	x	x			x		x
sklearn [84]			x				
statistics					x		
string			x				
traceback					x		
sys	x	x					
uuid				x			
werkzeug					x		

Table 8.1: Module usage per Main Server Block module

8.3.1 Web Server

The Web Server is the back-end for the Client, allowing him selected access to the Database. To build the Server, the authors chose to use the Flask Framework for Python 3.10 due to our familiarity with the software and its ease of use, efficient development process and expandability [85].

In our chosen implementation, the Web Server is the center block to the entire Main Server, it makes use of Python's `import` functionality to include the necessary core libraries and the developed modules:

- Extractor
- Preprocessor
- Mapper
- Database Connector

.

The directory tree implemented in the Web Server is presented in Figure 8.4, with variables highlighted in *italic text*. We opted to route everything related to the Extractor and Preprocessor under a single directory (/extractor) for brevity, as there is no apparent benefit from separating the results of these two modules' outputs.

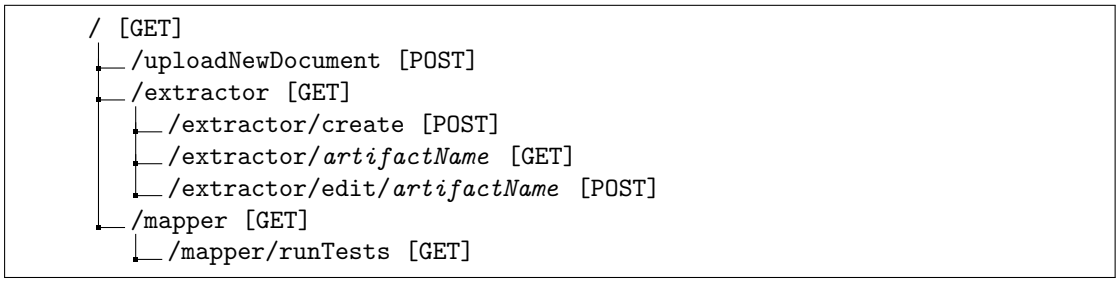

Figure 8.4: Directory structure of the Web Server

8.3.2 Templating Engine

Templating is a strategy that has gained a lot of popularity in the Web Development industry [86]. Templating makes use of a Templating Engine that takes as input a template file and, optionally, a set of variables. The output is a compiled version of the template with the variables replacing

the placeholders pre-set into the template. We opted to use the Jinja2 Templating Engine since it is built with our Web Server framework, Flask, in mind and works natively right out of the box.

A main template was built that includes:

- **Header -** Includes all of the styling (CSS) and some scripting (JS) files required for every page's display. A navigation bar is also included that allows transitioning between the extractor and mapper
- **Body -** Includes the HTML for the page to be rendered with the given variables
- **Footer -** Includes the remainder of the necessary scripting files (that must load after the page does)

8.3.3 Database Connector

Connections with the MySQL server are managed through the Database Connector module. The connector provides not only the interface with which to connect to the database, but also an abstraction for all the queries required by the Webserver.

A .env (environment) file is read when the module is loaded, which provides the necessary environment variables to connect with the database, namely, the username, password, host name and database name. The .env file allows any user to configure which database to use, without requiring any knowledge of implementation or even programming expertise.

8.3.4 Extractor

The Extractor is the first processing module in the pipeline. This module takes as input an ISO document in the PDF file format and the following user-supplied parameters:

- **Document Name (string) -** File name of the ISO document
- **Start Page (integer) -** Starting page for the Extractor to parse
- **End Page (integer) -** Last page for the Extractor to parse
- **Start Number (integer) -** First section for the Extractor to parse
- **Force Compute (boolean) -** "True" values force the extractor to run, ignoring any cached files

- **Input Path (string, optional) -** Path to the PDF file of the ISO document. If omitted, the resources folder is used by default
- **Output Path (string, optional) -** Path to the output folder. If omitted, the ”outputs” folder is used by default

The Extractor's main job is to extract the contexts from the PDF file supplied and convert them into a machine readable format so that it can be re-organized and outputted into the desired structure.

In order to perform the PDF extraction, we used an open-source module, pdfplumber [83]. This module gave us the full text of the PDF in a list of strings, which we merged into a single string. The string is passed on to a function that will utilize Regular Expressions (regex) to find the text related to its Sections and Subsections.

To facilitate the processing, an abstraction class was created for Sections and Subsections, StructuralElement, which encompasses both. A StructuralElement has an identifier, a title and its content. The identifier is the section or subsection numbering, the title is the section or subsection title and the content is all the text that the section or subsection encompasses.

StructuralElements are stored in a StructuralElementList, which is an extended Python list. This custom list type was designed to allow us to query by identifier or to filter by only section or subsection.

The StructuralElementsList gathered with this module is written to the predefined output file and returned from the main function.

8.3.5 Preprocessor

The Preprocessor was designed to take as input, the Extractor's output. However, the module can be used on its own with a StructuredElementsList originating from any source. This module's main job is to parse the StructuredElements' content to generate a mapping of the identifiers to a list of atomic controls. In order to find controls within each section and subsection's text, we have developed an an algorithm that utilizes Natural Language Toolkit (NLTK) [80] and custom-built regular expressions to identify controls.

Our algorithm starts by loading each StructuralElement's content into NLTK's `sent_tokenize` that separates the sentences and, by consence, the non-atomic controls that have yet to be processed. These are stored into an intermediate mapping of identifier to list of non-atomic controls.

At this stage, we pipe the list from each intermediate mapping into another function that parses each control to try to find enumerations and expands them into atomic controls by taking their header and applying it to each enumerated sentence. The lists are returned back to generate the final mapping, from identifier to a list of atomic controls. This mapping is returned back from the main function. Externally, we refer to this mapping as Preprocessor Artifact.

8.3.6 Mapper

The Mapper's goal is to generate a Mapping Artifact out of the two inputted Preprocessor Artifacts. Mapping Artifacts artifacts are generated, through IR techniques, in a four step sequential process composed by Tokenization, Normalization, TF-IDF and Comparison.

We developed and tested this module with a focus on English, although it could be utilized in other languages with minor configuration changes. As such, we follow the standard English language models and stopwords conventions.

8.3.6.A Tokenization

Classically, Tokenization is the first step to take on every NLP pipeline [87]. This step expands on the sentence tokenization and processing performed by the Preprocessor, taking the generated sentences and generating a list of tokens for them.

In order to generate a list of tokens for every sentences, we iterated over the list of controls and applied NLTK's [80] `word_tokenize` algorithm.

8.3.6.B Normalization

The first normalization step we took was to remove the English stopwords from every token list. This step is shown to improve the accuracy of IR techniques like sentiment analysis [88], which uses TF-IDF, the same algorithm we use.

Secondly, we piped the token lists to a function that modifies the corpus of the artifact, the set of all words, by transforming them through one of these methods:

- **Raw -** No changes are made to the corpus.
- **Stemming -** The function performs a stemming operation on all tokens in the corpus
- **Lemmatization -** The function performs a lemmatization operations on all tokens in the corpus

Out of the three methods, we have set Lemmatization as the default behaviour for the system, which is easily changeable with just some small changes to the configuration file. The impact on Precision and Recall is shown to be relatively small [89], but Lemmatization reduces the corpus size considerably, leading to lowered computational times in future steps.

8.3.6.C TF-IDF Comparisons

To create a Mapping Artifact, we have to compare the two normalized Preprocessor Artifacts, which is the goal of this step. We have implemented two algorithms for this effect: TF-IDF and Cosine Similarity.

TF is a vector containing the number of occurrences of a corpus' terms, for a given document. In this document, we define ”term” as processed token and ”document” as atomic control. The goal of TF is to measure which terms are relevant in a given document.

At last, TF-IDF is the simple multiplication of TF by IDF for a given term, in a given document from a given Artifact. This gives us the relevance of a token in relation to its context. By calculating these values to all of the controls in either of the two Preprocessor Artifacts, we gain a base we can use to compare them.

In order to perform the mapping, we needed to determine which controls were the most similar. With this goal in mind, we implemented a function to calculate the cosine similarity between the TF-IDF values of the two artifacts.

To compute the cosine similarity(csim), we first declared a new N x M zero matrix, with N equaling the number of controls for the Preprocessor Artifact (pa) 1 and M equaling the same for Preprocessor Artifact 2. The matrix is then filled out as described in Section 7.3.1, using sklearn's cosine similarity function [84].

The result is a similarity matrix, with higher numbers in a cell leading to a higher likelyhood of a mapping between the denoted controls.

8.3.6.D Neural Network Comparisons

The first implementation choice the authors have made, was the decision of which of the many sentence transformers to use. We have selected two, which have been trained with two distinct sets of documents: bert-base-nli-mean-tokens [90] and legal-bert-base-uncased [91].

We have opted to experiment with a general case sentence transformer, BERT, since it should work for any given input with reasonable accuracy. This model should provide a baseline

of what is possible to achieve.

Legal BERT, as the name would suggest, was trained under a large collection of legal documents. We assess that, due to the proximity of the compliance and legal fields, there is a fair possibility that comparisons made with a model specialized for law be as good, or better than those made with a generalized model.

8.3.6.E Merging the Comparison Results

Our proposal suggested that merging the results requires a fair amount of data processing and a form of normalization.

We started by doing the grouping part of the selection through a MySQL GROUP BY (clause1, clause2) and selecting the average of each similarity score, individually. This results in a mapping of the form $(clause1, clause2) : [simScoreNLP, simScoreAI]$ which we then passed onto a Pandas DataFrame [82].

We opted to introduce the StandardScaler implementation of the sklearn module [84]. This module allows us to, with few lines, normalize the two similarity arrays by zscore. However, it changes the scale from [0, 1], so we must set it back in the following lines. A summary of the normalization process is shown in Algorithm 8.1.

Algorithm 8.1: Summary of the normalization process

begin

$similarityValues \leftarrow df[[\text{"}value\text{"}, \text{"}valueAI\text{"}]].value$

$scaler \leftarrow StandardScaler()$

$normalizedData \leftarrow scaler.fit_transform(data)$

$df[\text{"}valueNormalized\text{"}], df[\text{"}valueAINormalized\text{"}] \longleftarrow normalizedData[:, 0], normalizedData[:, 1]$

$df[\text{"}valueNormalized\text{"}] \leftarrow setMinMax(df[\text{"}valueNormalized\text{"}])$

$df[\text{"}valueAINormalized\text{"}] \leftarrow setMinMax(df[\text{"}valueAINormalized\text{"}])$

$df[\text{"}result\text{"}] \leftarrow df[\text{"}valueNormalized\text{"}] * W1 + df[\text{"}valueAINormalized\text{"}] * W2$

8.3.7 Observing the results

The mapper presents the output artifact in the form of a mapping table. Each cell of the table is color coded based on the quantiles defined in the configuration file. The researcher can swap

out which view he desires: the weighted (combined) NLP and AI model or either of them separately. Example usage can be found in Figure 8.5. Outputs can be presented in continuous or discretized to values between 0 to 4, by selecting the appropriate mode (continuous or discrete).

8.4 MySQL Database

The MySQL Database acts as a data warehouse for all the persistent data in our system. We have encased it inside a docker container and volume, allowing for portability of the data within it. The database architecture can be found in 8.6

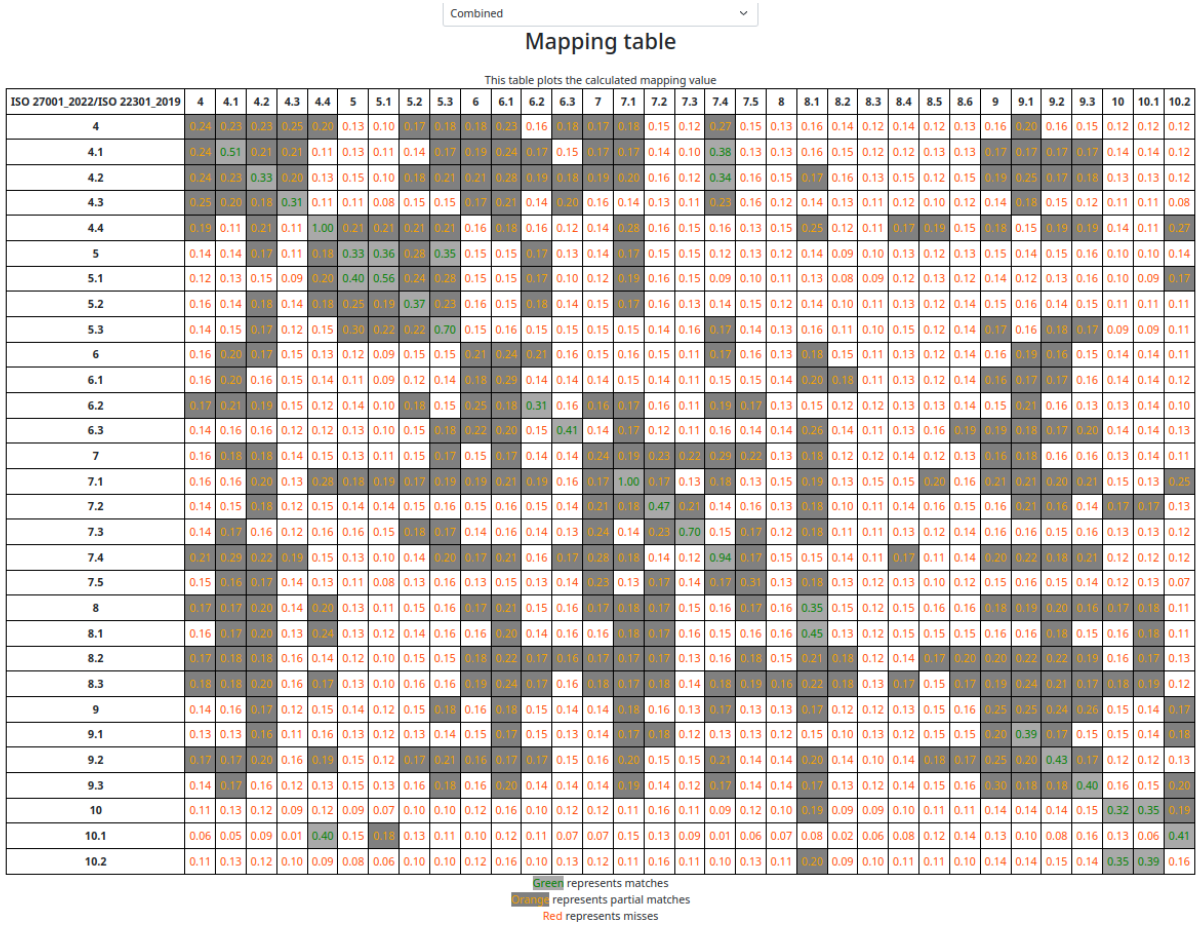

Figure 8.5: Mapping table for ISO 27001 2022 and ISO 22301 2019 [2, 3]

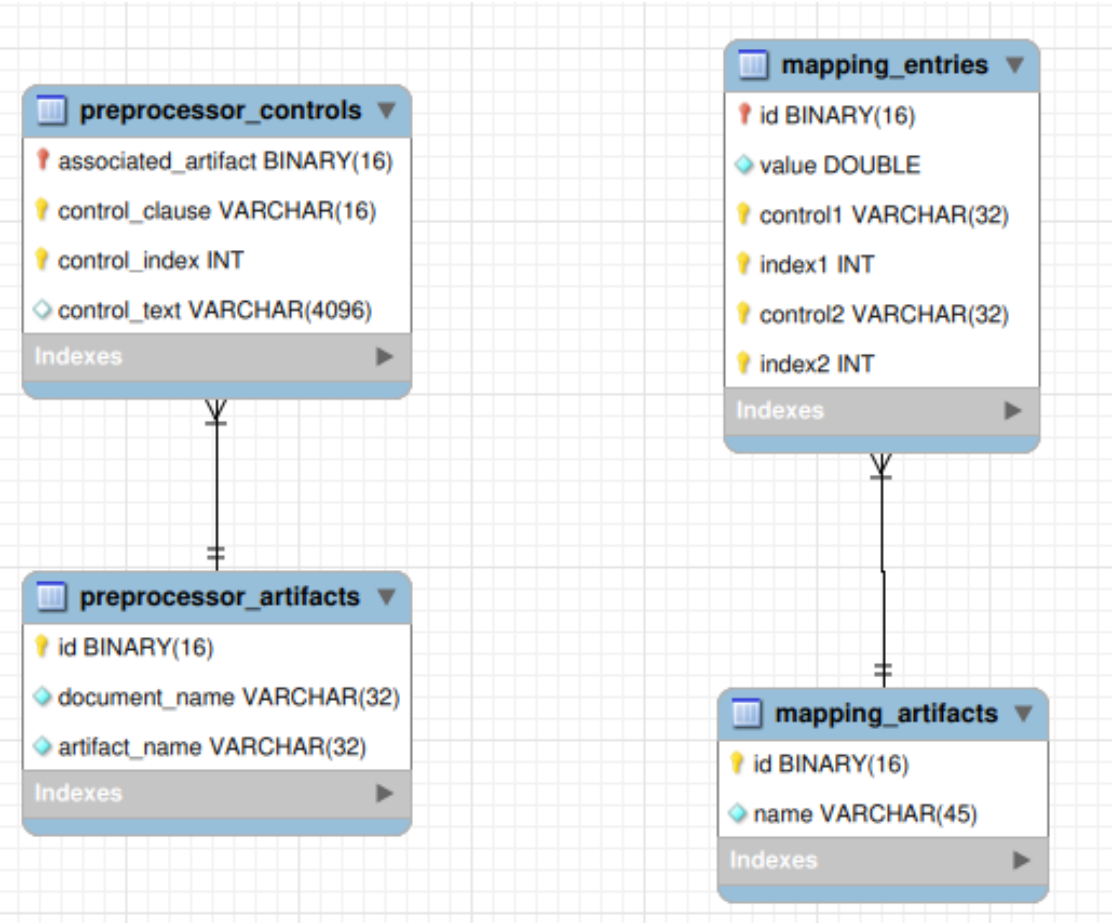

Figure 8.6: Database structure

9

Evaluation

Contents

9.1 Introduction

We have opted to evaluate this tool by comparing the mapping tables produced by the tool to those created manually. In such a model, we would be utilizing the mappings developed manually by humans as the ”golden standard”. For our approach to be viable, the mappings with which to compare should be the result of peer-reviewed studies, or at least, have some form of specialist input.

9.2 Evaluation methodology

We evaluated the tool utilizing two mappings: ISO 27001:2005 [4] with ISO 20000:2005 [5]and ISO 27001:2022 [2] with ISO 22301:2019 [3]. Both mappings follow the same format: a mapping table, with each cell being a unique pair of sections relating to two distinct compliance mechanisms.

The mapping table generated for the standards ISO 27001:2005 [4] and ISO 20000 [5] is compared against the one created manually by Pardo et al. [1]. This mapping table is a partial mapping in which each row or column has at least a non-zero value.

In regards to the mapping table generated for the standards ISO 27001:2022 and ISO 22301, we have taken a different approach. We have manually mapped between these two standards. These results were then validated with a survey that we handed out to specialists. The survey gathered 10 responses.

Having presented the input artifacts for the testing, it is also important to present the metrics we utilize to evaluate how well the tool performs. These metrics will depend on which ”mode” the tool is currently operating in: discrete or continuous. A summary of where we used each method is presented in Table 9.1.

Metric	Discrete mode	Continuous mode
R^2	X	X
Average error	X	X
F1-Score	X	
Spearman’s RCC		X

Table 9.1: Summary of the testing methods used per mode

9.2.1 Measurements common to both modes

For the common measurements, we have selected two: the R^2 value, which measures how good a fit does our model present before the test data; and the average error from the test data, given by the formula $\frac{1}{I*J}\sum\sum_{i\in I, j\in J}|generated(i,j) - test(i,j)|$, assuming that I and J are the number of rows and columns, respectively, `generated` is the value generated by the tool and `test` is the value pulled from the artifact.

Regarding the average error from the test data: this value always produces a value of [0, 4], in discrete mode, or [0, 1] in continuous mode. Lower values mean a lower average error.

9.2.2 Discrete measurements

The only specific measurement we picked here was the F1-score, given by the harmonic mean of the weighed Precision and Recall. In specific, we aim to target two different cases with this method: the F1-score for the non-zero labels and the specific F1-score for the zero label.

We opted to separate the two F1 measures due to the context in which the tool is meant to be used. If the main use a researcher is giving our tool is to assist them in selecting which instances are most likely to present a mapping between clauses, then it would be of the utmost importance to immediately exclude all the clauses that have nothing in common, which would be the majority, in most cases.

9.2.3 Continuous measurements

Once again, we present a single specific measurement in this subsection: the Spearman's Rank Correlation Coefficient (Spearman's RCC). We hope to determine how much agreement there is between the ranking performed by the tool and one carried out manually, by a human.

9.3 Evaluation results

We have ran both artifacts through ten variations of the testing suite, each one presents a unique combination of the weights given to the values generated by the AI model (AI %) and which model was used. A summary of the testing results is present in Tables 9.3 and 9.2. The best results were highlighted in **bold**.

Model	AI %	Avg Error (D)	$\mathbf{R^2(D)}$	Avg Error (C)	$\mathbf{R^2(C)}$	Zero F1	Non Zero F1	Spearman
None	0	**0,06**	**0,68**	0,11	-0,45	**0,99**	**0,34**	**0,62**
bert-base-nli-mean-tokens	25	**0,06**	0,65	**0,03**	**0,50**	**0,99**	0,31	0,38
bert-base-nli-mean-tokens	50	0,07	0,53	0,12	-0,97	0,98	0,31	0,28
bert-base-nli-mean-tokens	75	0,10	0,30	0,23	-5,08	0,98	0,10	0,26
bert-base-nli-mean-tokens	100	0,12	0,01	0,34	-12,00	0,97	0,14	0,22
nlpaueb/legal-bert-base-uncased	25	**0,06**	0,69	0,06	0,09	**0,99**	0,31	0,30
nlpaueb/legal-bert-base-uncased	50	0,07	0,56	0,05	0,20	0,98	0,24	0,30
nlpaueb/legal-bert-base-uncased	75	0,09	0,30	0,11	-0,88	0,98	0,24	0,27
nlpaueb/legal-bert-base-uncased	100	0,12	-0,06	0,17	-3,06	0,97	0,17	0,22

Table 9.2: Testing results for ISO 27001 2022-ISO 22301 2019 [2, 3]

Model	AI %	Avg Error (D)	$\mathbf{R^2(D)}$	Avg Error (C)	$\mathbf{R^2(C)}$	Zero F1	Non Zero F1	Spearman
None	0	0,98	-0,65	0,22	-0,62	**0,72**	0,23	**0,17**
bert-base-nli-mean-tokens	25	**0,97**	**-0,63**	**0,18**	-0,26	**0,72**	**0,26**	0,12
bert-base-nli-mean-tokens	50	1,00	-0,71	0,21	**-0,16**	0,71	**0,26**	0,10
bert-base-nli-mean-tokens	75	1,06	-0,79	0,24	-0,33	0,70	0,19	0,13
bert-base-nli-mean-tokens	100	1,11	-0,98	0,28	-0,78	0,67	0,23	0,09
nlpaueb/legal-bert-base-uncased	25	1,05	-0,82	**0,18**	-0,32	0,71	0,23	0,13
nlpaueb/legal-bert-base-uncased	50	1,11	-0,88	0,21	-0,24	0,67	0,16	0,04
nlpaueb/legal-bert-base-uncased	75	1,16	-1,07	0,23	-0,32	0,65	0,21	-0,02
nlpaueb/legal-bert-base-uncased	100	1,17	-1,07	0,27	-0,69	0,65	0,19	-0,09

Table 9.3: Testing results for ISO 27001 2005-ISO 20000 2005 [4, 5]

9.4 Evaluation Analysis

In this subsection, we will analyze and discuss the previously presented results in more depth. Firstly, we will examine the overall results and evaluate the effectiveness of the tool in performing its function. We explore whether the incorporation of neural networks has contributed to any improvements in the results and, if so, determine the optimal weights assigned to them.

9.4.1 Analysing the results as a whole

Our primary goal was to achieve, at the very least, a good classification on which controls can be excluded from a mapping, which would be the overwhelming majority of the total number of control pairs. To evaluate performance in this setting, we can find insight from four metrics: the average error, the R^2, the Spearman coefficient and the F1 Score for the zero label in specific. The former three measures are more generalized, while the latter predicts, specifically, whether or not the tool is accurately predicting a binary ”is this pair of controls similar at all?”

We first observe a large disparity between the testing results, which we can explain due to the nature of the two artifacts and the difference in F1-Scores. The tool appears to be extremely effective at predicting whether or not a value has any use, as the F1-Scores for the zero label are extremely high for both artifacts. This fact, combined with the knowledge that the ISO 27001_2005 - ISO 20000_2005 [4, 5] mapping presents only a partial mapping, with much less

0 values as a whole, leads to the whole set of results being worse.

Secondly, we note that the average error is low and that the Spearman coefficient values are positive, meaning that there is a monotonic relationship, a correlation, between the test values. Both of these values indicate that the tool is better than the average case. However, the partial mapping shows significantly worse results, a conclusion which is furthered by the R^2 values, which are overwhelmingly positive in the second artifact, but negative on the first. This means a large part of the variance between test and predicted results on the first artifact is explained by our tool, while no such conclusion can be taken on the second.

9.4.2 Analysing the effectiveness of the selected neural network models

In order to take conclusions on the effectiveness of the selected neural networks, we will put into the spotlight the generalized metrics: average error, R^2 and the Spearman coefficient.

We observe that the best values for both tables, across all of these metrics, are obtained when using either 0% AI weights, 25% or 50%. The use of higher weights in this interval tends to be more effective at lowering the average error in continuous mode, as compared to the lower weights which appear to be best in discrete mode. The one exception to this rule is the Spearman coefficient, which proves to be highest at 0% AI usage.

The differences between the two chosen models were small. However, the generalized BERT model appears to be slightly more precise, on average, across all metrics. We see no indication that utilizing the Legalbert model would be any better in practice for our specific use case. We assess that the best way to parametrize the tool, with this information in mind, would be to always use the ”bert-base-nli-mean-tokens” model at weights of 0 to 25%, as these proved to be the most effective in this phase.

The present data leads us to conclude that the use of the chosen AI models to assist in the mapping process is no better than the classical NLP techniques, at least not standalone. We hypothesize that this fact could be owed to the way the two models chosen were trained. Both BERT and LegalBERT were fed their own datasets and evaluated the performance of the results using cosine similarity, the same metric we use for NLP text comparison, leading to fairly similar results.

10

Conclusion

Implementing compliance mechanisms, such as standards, policy, or legislation, offer organizations numerous benefits. However, blindly implementing them can lead to, among other problems, needlessly duplicated work and increased costs. The main solution to these problems is to carry out a mapping study between the mechanisms that the organisation aims to see implemented and the ones already in use within the organization.

The mapping study provides a clear overview of shared requirements across documents, but it comes with the significant time investment required to produce an artifact. Through the application of the Design Science Research Methodology and a Systematic Literature Review, we identified this time-consuming process as the primary research problem.

To tackle this research problem, we designed, proposed, developed, and evaluated a software tool that automates parts of the homogenization and mapping processes. Our tool aims to extract atomic requirements from a pair of ISO standards and determine the likelihood of their presence in both standards. When not utilizing ISO standards, the first step can be bypassed by manually extracting the requirements into a spreadsheet and loading it into the tool.

Our developed tool utilizes two methods to perform the comparisons between atomic controls: one is presented as the ”classical” way to do Short Text Similarity, through cosine similarities of TF-IDF vectors; while the other utilizes recent developments in Artificial Intelligence, by applying two pre-trained neural network models to perform the cosine similarities. Out of the two models we picked, we chose one that has a generalized use case, the DistilBERT model [56,90] and another, LegalBERT [91], that was specifically trained on legal documents, since the field of compliance is closely linked with the legal field.

To evaluate the tool's performance, we compared its automatically generated results against two manually created mappings considered as the ”golden standard.” These mappings comprised mapping tables derived from the literature [1] and an in-house table evaluated by specialists through a survey.

The evaluation demonstrates that the tool performs reasonably well, with low average errors when using the default parameters. Particularly, it shows excellent results in identifying controls that can be excluded from the mapping, which constitute the majority of control pairs.

The tool and the research conducted during its planning and development constitute our main contributions to the industry. We have submitted two papers for peer review: one focusing on the SLR, which advances the State of the Art in the field, and another presenting the tool comprehensively, from proposal to evaluation.

However, it is essential to acknowledge the limitations of our work. We have solely evaluated

the tool using ISO standards and applied relatively similar AI models derived from BERT [56, 90, 91].

For Future Work, we propose that the testing could be expanded to other types of compliance mechanisms, so that the tool's efficacy can be studied when applied to them. Interesting examples could include organizational policy, other standards like the COBIT Framework [14] or legislation (such as the GDPR [13]). We also suggest that other AI models could be studied, preferably utilizing a different basis for their training, possibly utilizing novel techniques such as Word Mover's Distances (WMD) [92].

Bibliography

[1] C. Pardo, F. J. Pino, and F. Garcia, “Towards an integrated management system (ims), harmonizing the iso/iec 27001 and iso/iec 20000-2 standards,” *International Journal of Software Engineering and Its Applications*, vol. 10, no. 9, pp. 217–230, 2016.

[2] ISO Central Secretary, “Information security management,” International Organization for Standardization, Geneva, CH, Standard, 2022.

[3] ——, “Security and resilience,” International Organization for Standardization, Geneva, CH, Standard, 2019.

[4] ——, “Information security management,” International Organization for Standardization, Geneva, CH, Standard, 2005.

[5] ——, “Service management system requirements,” International Organization for Standardization, Geneva, CH, Standard, 2005.

[6] G. A. Feltham, “The value of information,” *The accounting review*, vol. 43, no. 4, pp. 684–696, 1968.

[7] M. Gerber and R. Von Solms, “Management of risk in the information age,” *Computers & security*, vol. 24, no. 1, pp. 16–30, 2005.

[8] M. M. Silva, A. P. H. de Gusmão, T. Poleto, L. C. e Silva, and A. P. C. S. Costa, “A multidimensional approach to information security risk management using fmea and fuzzy theory,” *International Journal of Information Management*, vol. 34, no. 6, pp. 733–740, 2014.

[9] K. Kaur, I. Gupta, and A. K. Singh, “Data leakage prevention: e-mail protection via gateway,” in *Journal of Physics: Conference Series*, vol. 933, no. 1. IOP Publishing, 2017, p. 012013.

[10] B. Morrow, “Byod security challenges: control and protect your most sensitive data,” *Network Security*, vol. 2012, no. 12, pp. 5–8, 2012.

[11] H. Tohidi, “The role of risk management in it systems of organizations,” *Procedia Computer Science*, vol. 3, pp. 881–887, 2011.

[12] K. Julisch, “Security compliance: the next frontier in security research,” in *Proceedings of the 2008 New Security Paradigms Workshop*, 2008, pp. 71–74.

[13] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[14] IT Governance Institute, Ed., *CobiT 4.1: Framework, Control Objectives, Management Guidelines, Maturity Models*. Rolling Meadows: IT Governance Institute, 2007.

[15] R. Sheikhpour and N. Modiri, “An approach to map cobit processes to iso/iec 27001 information security management controls,” *International Journal of Security and Its Applications*, vol. 6, no. 2, pp. 13–28, 2012.

[16] A. Simon, S. Karapetrovic, and M. Casadesús, “Difficulties and benefits of integrated management systems,” *Industrial Management & Data Systems*, 2012.

[17] M. Yasin, A. A. Arman, I. J. M. Edward, and W. Shalannanda, “Designing information security governance recommendations and roadmap using cobit 2019 framework and iso 27001: 2013 (case study ditreskrimsus polda xyz),” in *2020 14th International Conference on Telecommunication Systems, Services, and Applications (TSSA*. IEEE, 2020, pp. 1–5.

[18] B. Barafort, A.-L. Mesquida, and A. Mas, “Integrating risk management in it settings from iso standards and management systems perspectives,” *Computer Standards and Interfaces*, vol. 54, pp. 176–185, 2017.

[19] K. Beckers, S. Hofbauer, G. Quirchmayr, and C. C. Wills, “A method for re-using existing itil processes for creating an iso 27001 isms process applied to a high availability video conferencing cloud scenario,” in *Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013. Proceedings 8*. Springer, 2013, pp. 224–239.

[20] C. Magnusson and S.-C. Chou, “Risk and compliance management framework for outsourced global software development,” *Proceedings - 5th International Conference on Global Software Engineering, ICGSE 2010*, pp. 228–233, 2010.

[21] O. Mangin, B. Barafort, P. Heymans, and E. Dubois, “Designing a process reference model for information security management systems,” in *Software Process Improvement and Capability Determination: 12th International Conference, SPICE 2012, Palma, Spain, May 29-31, 2012. Proceedings 12*. Springer, 2012, pp. 129–140.

[22] A.-L. Mesquida, A. Mas, T. S. Feliu, and M. Arcilla, “Min-its: A framework for integration of it management standards in mature environments,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 24, pp. 887–908, 2014.

[23] H. Muzaimi, B. C. Chew, and S. R. Hamid, “Integrated management system: The integration of iso 9001, iso 14001, ohsas 18001 and iso 31000,” in *AIP conference proceedings*, vol. 1818, no. 1. AIP Publishing LLC, 2017, p. 020034.

[24] C. Pardo, F. J. Pino, and F. Garcia, “Towards an integrated management system (ims), harmonizing the iso/iec 27001 and iso/iec 20000-2 standards,” *International Journal of Software Engineering and its Applications*, vol. 10, pp. 217–230, 2016.

[25] E. Pretorius and B. von Solms, “Information security governance using iso 17799 and cobit,” in *Integrity and Internal Control in Information Systems VI: IFIP TC11/WG11. 5 Sixth Working Conference on Integrity and Internal Control in Information Systems (IICIS) 13–14 November 2003, Lausanne, Switzerland*. Springer, 2004, pp. 107–113.

[26] H. Rahmani, A. Sami, and A. Khalili, “Cip-uqim: A unified model for quality improvement in software sme’s based on cmmi level 2 and 3,” *Information and Software Technology*, vol. 71, pp. 27–57, 2016.

[27] S. Ramanauskaite, D. Olifer, N. Goranin, and A. Čenys, “Security ontology for adaptive mapping of security standards,” *International Journal of Computers, Communications and Control*, vol. 8, pp. 878–890, 2013.

[28] K. Ruamchat, N. Thawesaengskulthai, and C. Pongpanich, “Development of quality management system under iso 9001:2015 and joint inspection group (jig) for aviation fuelling service,” *Management and Production Engineering Review*, vol. 8, pp. 50–59, 2017.

[29] B. V. Solms, “Information security governance: Cobit or iso 17799 or both?” *Computers and Security*, vol. 24, pp. 99–104, 2005.

[30] A. Hevner, S. Chatterjee, A. Hevner, and S. Chatterjee, “Design science research in information systems,” *Design research in information systems: theory and practice*, pp. 9–22, 2010.

[31] B. Kitchenham, “Procedures for performing systematic reviews,” *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[32] R. Almeida, R. Lourinho, M. M. da Silva, and R. Pereira, “A model for assessing cobit 5 and iso 27001 simultaneously,” in *2018 IEEE 20th Conference on Business Informatics (CBI)*, vol. 1. IEEE, 2018, pp. 60–69.

[33] L. Rittenberg, F. Martens, C. of Sponsoring Organizations of the Treadway Commission *et al.*, “Enterprise risk management: understanding and communicating risk appetite,” 2012.

[34] M. Power, “The risk management of nothing,” *Accounting, organizations and society*, vol. 34, no. 6-7, pp. 849–855, 2009.

[35] ISO Central Secretary, “Risk management,” International Organization for Standardization, Geneva, CH, Standard, 2018.

[36] M. Gerber and R. Von Solms, “Information security requirements–interpreting the legal aspects,” *Computers & Security*, vol. 27, no. 5-6, pp. 124–135, 2008.

[37] ISO Central Secretary, “Information technology — security techniques — code of practice for information security controls,” International Organization for Standardization, Geneva, CH, Standard, 2013.

[38] C. Pardo, F. J. Pino, F. García, M. Piattini, and M. T. Baldassarre, “An ontology for the harmonization of multiple standards and models,” *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 48–59, 2012.

[39] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, “What are ontologies, and why do we need them?” *IEEE Intelligent Systems and their applications*, vol. 14, no. 1, pp. 20–26, 1999.

[40] J. Siviy, P. Kirwan, L. Marino, and J. Morley, “The value of harmonizing multiple improvement technologies: a process improvement professional’s view,” CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, Tech. Rep., 2008.

[41] C. Pardo, F. J. Pino, F. García, M. Piattini, and M. T. Baldassarre, “A process for driving the harmonization of models,” in *Proceedings of the 11th International Conference on Product Focused Software*, 2010, pp. 51–54.

[42] ——, “An ontology for the harmonization of multiple standards and models,” *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 48–59, 2012.

[43] L. Ibrahim and A. Pyster, “A single model for process improvement,” *IT professional*, vol. 6, no. 3, pp. 43–49, 2004.

[44] C. Yoo, J. Yoon, B. Lee, C. Lee, J. Lee, S. Hyun, and C. Wu, “A unified model for the implementation of both iso 9001: 2000 and cmmi by iso-certified organizations,” *Journal of Systems and Software*, vol. 79, no. 7, pp. 954–961, 2006.

[45] C. J. P. Calvache, F. J. Pino, F. García, and M. Piattini, “Homogenization of models to support multi-model processes in improvement environments.” in *ICSOFT (1)*, 2009, pp. 151–156.

[46] Ž. Turk, “Construction informatics: Definition and ontology,” *Advanced engineering informatics*, vol. 20, no. 2, pp. 187–199, 2006.

[47] A. Mas, A. L. Mesquida, E. Amengual, and B. Fluxà, “Iso/iec 15504 best practices to facilitate iso/iec 27000 implementation,” in *International Conference on Evaluation of Novel Approaches to Software Engineering*, vol. 2. SciTePress, 2010, pp. 192–198.

[48] M. Koubarakis, *Telos*. Boston, MA: Springer US, 2009, pp. 2914–2920.

[49] ISO Central Secretary, “Information technology — process assessment — requirements for process reference, process assessment and maturity models,” International Organization for Standardization, Geneva, CH, Standard, 2015.

[50] K. Chowdhary and K. Chowdhary, “Natural language processing,” *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[51] V. Balakrishnan and E. Lloyd-Yemoh, “Stemming and lemmatization: A comparison of retrieval performances,” in *Proceedings of SCEI Seoul Conferences, 10-11 Apr 2014, Seoul, Korea*, 2014.

[52] A. Aizawa, “An information-theoretic perspective of tf–idf measures,” *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[53] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[54] P. P. Shinde and S. Shah, “A review of machine learning and deep learning applications,” in *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[57] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[58] R. Grishman, “Information extraction,” *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 8–15, 2015.

[59] A. K. Mahapatra and S. Biswas, “Inverted indexes: Types and techniques,” *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 4, p. 384, 2011.

[60] A. Mussmann, M. Brunner, and R. Breu, “Mapping the state of security standards mappings.” in *Wirtschaftsinformatik (zentrale tracks)*, 2020, pp. 1309–1324.

[61] N. K. Gunawan, R. B. Hadiprakoso, and H. Kabetta, “Comparative study between the integration of itil and iso/iec 27001 with the integration of cobit and iso/iec 27001,” in *IOP Conference Series: Materials Science and Engineering*, vol. 852, no. 1. IOP Publishing, 2020, p. 012128.

[62] R. Sandadi, *ITIL Foundation Reference Guide: Concepts, Use Case, Exam Guide*. Independently published, 2017.

[63] C. Pardo, F. J. Pino, F. García, M. Piattini Velthius, and M. T. Baldassarre, "Trends in harmonization of multiple reference models," in *Evaluation of Novel Approaches to Software Engineering: 5th International Conference, ENASE 2010, Athens, Greece, July 22-24, 2010, Revised Selected Papers 5*. Springer, 2011, pp. 61–73.

[64] D. W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review," *Soft Computing*, vol. 25, pp. 4699–4723, 2021.

[65] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A comparative study of two short text semantic similarity measures," in *Agent and Multi-Agent Systems: Technologies and Applications: Second KES International Symposium, KES-AMSTA 2008, Incheon, Korea, March 26-28, 2008. Proceedings 2*. Springer, 2008, pp. 172–181.

[66] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[67] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," Technical report, Ver. 2.3 EBSE Technical Report. EBSE, Tech. Rep., 2007.

[68] I. M. Lopes, T. Guarda, and P. Oliveira, "How iso 27001 can help achieve gdpr compliance," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2019, pp. 1–6.

[69] S. M. Castillo-Rojas, M. Casadesús, S. Karapetrovic, L. Coromina, I. Heras, and I. Martín, "Is implementing multiple management system standards a hindrance to innovation?" *Total Quality Management & Business Excellence*, vol. 23, no. 9-10, pp. 1075–1088, 2012.

[70] C. Vroom and R. Von Solms, "Towards information security behavioural compliance," *Computers & security*, vol. 23, no. 3, pp. 191–198, 2004.

[71] A. M. Picahaco, A. L. Mesquida, E. A. Alcover, and B. Fluxà, "Iso/iec 15504 best practices to facilitate iso/iec 27000 implementation." in *ENASE*, 2010, pp. 192–198.

[72] S. Fenz and T. Neubauer, "Ontology-based information security compliance determination and control selection on the example of iso 27002," *Information & Computer Security*, 2018.

[73] P. Achananuparp, X. Hu, and X. Shen, “The evaluation of sentence similarity measures,” in *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings 10*. Springer, 2008, pp. 305–316.

[74] I. Atoum, A. Otoom, and N. Kulathuramaiyer, “A comprehensive comparative study of word and sentence similarity measures,” *arXiv preprint arXiv:1610.04533*, 2016.

[75] H. Abdi, L. J. Williams *et al.*, “Normalizing data,” *Encyclopedia of research design*, vol. 1, 2010.

[76] theskumar, “python-dotenv,” https://github.com/theskumar/python-dotenv, 2022.

[77] pallets, “flask,” https://github.com/pallets/flask, 2022.

[78] ——, “jinja,” https://github.com/pallets/jinja, 2022.

[79] M. Widenius and D. Axmark, *MySQL reference manual: documentation from the source*. ” O’Reilly Media, Inc.”, 2002.

[80] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.

[81] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[82] T. pandas development team, “pandas-dev/pandas: Pandas,” Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[83] jsvine, “pdfplumber,” https://github.com/jsvine/pdfplumber, 2022.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[85] P. Lokhande, F. Aslam, N. Hawa, J. Munir, and M. Gulamgaus, “Efficient way of web development using python and flask,” *International Journal of Advanced Research in Computer Science*, 2015.

[86] D. Gibson, K. Punera, and A. Tomkins, “The volume and evolution of web page templates,” in *Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005, pp. 830–839.

[87] J. J. Webster and C. Kit, “Tokenization as the initial phase in nlp,” in *COLING 1992 volume 4: The 14th international conference on computational linguistics*, 1992.

[88] K. V. Ghag and K. Shah, “Comparative analysis of effect of stopwords removal on sentiment classification,” in *2015 international conference on computer, communication and control (IC4)*. IEEE, 2015, pp. 1–6.

[89] M. Toman, R. Tesar, and K. Jezek, “Influence of word normalization on text classification,” *Proceedings of InSciT*, vol. 4, pp. 354–358, 2006.

[90] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *ArXiv*, vol. abs/1910.01108, 2019.

[91] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904.

[92] V. Novotnỳ, E. F. Ayetiran, M. Štefánik, and P. Sojka, “Text classification with word embedding regularization and soft similarity measure,” *arXiv preprint arXiv:2003.05019*, 2020.