CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

A Technique for Compliance Identification for Information Security Policy Documents

by

Maham Tariq

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2021

My work is dedicated, firstly to the Almighty Allah, for blessing me with the opportunities, health and abilities to be where I am do what I did. After Allah, this study is wholeheartedly dedicated to my beloved parents, who have been my biggest inspiration, and specially to my support system, my father, for his constant moral, spiritual, emotional, and financial support.

CERTIFICATE OF APPROVAL

A Technique for Compliance Identification for Information Security Policy Documents

by

Maham Tariq

(MCS191052)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Arshad Islam	FAST-NUCES, Islamabad
(b)	Internal Examiner	Dr. M. Masroor Ahmed	CUST, Islamabad
(c)	Supervisor	Dr. Qamar Mahmood	CUST, Islamabad

Dr. Qamar Mahmood
Thesis Supervisor
May, 2021

Dr. Nayyer Masood
Head
Dept. of Computer Science
May, 2021

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
Mayl, 2021

Author's Declaration

I, **Maham Tariq** hereby state that my MS thesis titled "**A Technique for Compliance Identification for Information Security Policy Documents**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Maham Tariq)

Registration No: MCS191052

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Technique for Compliance Identification for Information Security Policy Documents**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Maham Tariq)
Registration No: MCS191052

Acknowledgement

If you are grateful, I would certainly give you more; and if you are ungrateful, My chastisement is truly severe. (14:7) Alhamdulillah for everything Allah has blessed me with that helped me reach here.

I would like to show my gratitude to my supervisor, who shared his words of advice and encouragement to this study and helped me through every difficulty regarding this study.

I am thankful to my mother for pushing me for higher studies and my father for making everything I wish for possible for me.

Lastly, I would like to thank my sister and my friends for always being there when I needed help and support.

(Maham Tariq)

Abstract

The amount of data constantly being created is increasing with time; hence, it is becoming more and more critical to protect the data from security mishaps. Every organization needs a set of rules to protect its information and assets from internet and external security breaches. Such rules are usually stated in a security policy document. This document contains information about the security mechanisms and technologies being implemented and also explains the roles and responsibilities of every concerned employee. Information Security Policy Documents are receiving great attention from researchers since the early 2000s. Although security policy documents are the focus point of many recent research studies but there is very little content on making the task easier. The very few available solutions are either too complex and expensive or not very abstract. No concrete study has been found that suggests any technique to find compliance of information security policy documents to a standard template. In this study a technique is proposed which identifies the compliance of any given Information Security Policy Document with the standard template and calculate a compliance score which will help identify the degree of deviation from the standard document. Data is collected from the web resources of different healthcare organizations. The techniques used in this experiment are Cosine Similarity Measure, Jaccard Similarity Measure and String Similarity Measure. The final result is the weighted sum of these techniques. The results are evaluated with the help of standard evaluation measures like accuracy, precision, recall, f-measure. The results from user-based evaluation are considered as gold standard. The scores of the proposed technique came out to be similar to scores of user-based evaluations. The proposed technique is found to be 66% accurate. This study opens doors for future research in different domains. Multiple combination of similarity techniques can be applied and tested.

Contents

List of Figures

List of Tables

Abbreviations

Acronym	**W**hat (it) **S**tands **F**or
ISPD	Information Security Policy Document
NLTK	The Natural Language Toolkit is a package which contains many methods and libraries for statistical and lexical analysis of natural language
ISO	Organisation internationale de normalization - International Organization for Standardization
NIST	National Institute of Standards and Technology
NLC	National Learning Consortium
Punkt	Sentence Tokenizer found in nltk package

Chapter 1

Introduction

1.1 Background

Every organization needs a set of rules to protect its information and assets from internet and external security breaches. Such rules are usually stated in a security policy document. This document contains information about the security mechanisms and technologies being implemented and also explains the roles and responsibilities of every concerned employee. Some organizations prepare their information security policy documents themselves while some organizations follow the international security standards.

Due to lack of knowledge most of the time these documents contain loopholes. A single gap in security policy can cause great damage so it is very critical for the security policy to be complete and free of ambiguities. Development of security policy is a huge task in itself, reviewing it is another.

1.1.1 Compliance

Compliance is defined as the state of meeting some rules or standards. In this study, the word compliance refers to the degree of similarity of a document to another document chosen as standard.

1.1.2 International Security Standards

A standard is a published specification that establishes a common language, and contains a technical specification or other precise criteria and is designed to be used consistently, as a rule, a guideline, or a definition. Among other such organizations, ISO and NIST are considered as the most authentic in the field of information security as they are very comprehensive and refined. These organizations offer paid certifications and security policy compliance checks to organizations wanting to avoid security risks.

1.1.3 Developing an Information Security Policy Document

Whenever an organization develops an Information Security Policy Document it can either use an already existing policy, e.g., the International Security Standards or it can develop its own customized Policy Document from scratch and then evaluate it for any shortcomings. As the International Security Standards are expensive and might not fit perfectly with every organizations policies, most organizations opt for the second option. Not many tools are available to aid the process of evaluation of such Information Security Documents. This study will be helpful in solving this problem.

1.2 Motivation

Although security policy documents are the focus point of many recent research studies but there is very little content on making the task easier. Most of the studies emphasize on the development of a flawless security policy or on the management of the security policy. The theme of this research is to check the compliance of the security policy document with a standard document. Information Security policy developers will benefit from this work. Organizations will be able to check the legitimacy of their security policy documents very easily by using this technique. Generally, this study will help to improve security of organizations so it can be considered an addition to the field of information security technology.

1.3 Problem Statement

Information Security Policy Documents are presently the topic of interest of researchers but the in-depth analysis of the literature shows that previous research lacks any technique to solve the problem of non-compliance of information security policy documents with standards of security by identifying a compliance score. In this research, a technique is proposed which can solve this problem.

1.4 Research Problems

The above problem statement raises some research questions, which are stated below:

Q1: What are the various techniques being used for the calculation of document similarities and how they can be combined to build a security policy compliance finding model?

Q2: How document compliance finding model can be evaluated for better accuracy?

1.5 Research Methodology

The methodology of this research work is based on the experimental research method and it comprises of three main phases:

1. Exploratory
2. Implementation
3. Evaluation

1.5.1 Exploratory

This phase consists of a detailed study on the topic and review of the relevant literature found. This is done to identify the significance of the problem and find the shortcomings in already present solutions (if any).

1.5.2 Implementation

This phase is based on execution of the proposed solution. The proposed technique is implemented and compliance score of different documents is calculated.

1.5.3 Evaluation

The last phase of this research is the evaluation of the results computed from the proposed technique. The results computed from the proposed technique are discussed and compared. The purpose of this evaluation is to check the authenticity of the created tool.

1.6 Thesis Organization

In Introduction chapter, a brief overview about the topic is given and the problem is explained. The significance of the problem and research methodology is also discussed.

The chapter of Literature Review includes the findings of detailed literature survey that is performed to identify the implication of the problem. The content of chapter 2 answers the Research Question 1. The reviewed literature is discussed in detail and relevant information from the already existing literature is also added.

The Chapter 3 presents the step-by-step solution of the problem. A new technique is proposed that is expected to give best results as compared to techniques that are already being used, addressing the Research Question 2. The architecture of the designed technique is discussed along with the techniques being used to solve the problem.

The Chapter 4 discusses the final results calculate after implementation of the proposed solution show the performance of the proposed technique. These results are then evaluated in two steps, i.e., user-based testing and evaluation measures like precision, recall etc., thus solving the research question 3.

The findings of this study are concluded in the last chapter. The significance of the problem and its proposed solution is explained according to the results found in chapter 5. Future work is also suggested.

Chapter 2

Literature Review

2.1 Introduction

As the technology is progressing, the amount of data constantly being created is increasing; hence, the value of this data is also rising. It is becoming more and more critical to protect the data from security mishaps. Information Security Policy Documents are receiving great attention from researchers since the early 2000s. Research is being performed in different directions about the information security policy documents but after exhaustive research, it can be said that not much attention is being paid on the problem of information security policy compliance.

Most researchers focus on the development and implementation of sustainable information security policies.

Another area which is being heavily researched is the management of the information security policy document, effectively communicating the policies to employees and making them follow the policies.

2.2 Survey Questions

The detailed literature review is performed on the basis of the following questions:

Q1: Explore the available techniques to find compliance for security policy document and the Similarity Measures are being used in them.

Q2: Study the necessary policy requirements of an information security policy document.

Q3: Review the commonly used techniques for the calculation of similarity between documents.

Q4: Study the existing research on Information Security Policy Documents that suggest implementation of an automated tool for identification of policy compliance.

2.3 Surveyed Techniques

The literature analysis is focuses mainly on three topics. First; literature about compliance identification between documents, more importantly information security policy documents, second; studies about information security policy documents, their development, necessary requirements and the methods being used to validate these documents, third; the present literature about similarity techniques being used or studied to calculate document similarities.

In total, 40 research papers and 2 thesis documents have been studied, out of which 35 literature artifacts are found to be most relevant. Some of these are further discussed in detail in the following text.

2.3.1 Document Compliance Techniques

Buthelezi and Van [1] mentioned that ambiguity found in security policies can lead to non-compliance. Content Analysis performed on data collected from security policy documents from different organizations suggested that the policy writers should make a cognizant effort to express the policy statements explicitly with sufficient detail. It also proposed for future researchers to investigate methods for resolving ambiguities in information security documents with the help of software development.

Compliance identification with a standard template is not a new concept. It has been used in previous studies [2–4]. One of them proposed a compliance checking system where documents were compared with given templates. But it did not deal with security policy documents, it basically checked the compliance of IT services business contract documents to templates. It used vector space similarity measure for calculating document similarity. Two different techniques are used to compare contracts to templates to identify top candidate templates for more detailed analysis. Each technique depends on a term vector representation of a document. In one case, cosine similarity was used, whereas in the other case Latent Semantic Indexing was used for dimensionality reduction before applying cosine similarity. The prior discussed study is further enhanced in another study [5] where it is divided into three modules. The foremost measured the extent of compliance of original contracts to the standard templates. While the second module analyzed the compliance of those contracts which had adequate nonconformities and then the patterns of these nonconformities which were being repeatedly observed in the results were analyzed for every template. The last module analyzed the contracts which showed no compliance whatsoever and distinguished sets in the selected contracts such that items of every set must possess adequate similarity to each other to so that they can be considered for development of new templates for every set.

In this research, it is intended to calculate weighted similarity using multiple techniques. If the similarity between documents is based solely on matching phrases, and not single-terms at the same time, related documents could be judged as non-similar if they do not share enough phrases [6]. During the process of document clustering, a new similarity measure, i.e., Document Index Graph was introduced to calculate phrase-based similarity from a document by indexing the contents of document while preserving the sentence structure in the original document, cosine correlation similarity measure for the single term similarity. And then similarity based on the weights of both single word and phrase-based similarity measures was calculated. This study also proposed that the accuracy of similarity calculation between documents can be further improved by employing different similarity calculation strategies in future.

The present literature also contains study which suggest developing a computerized tool for information security policy documents [7]. This study aimed to elicit a set of requirements, anchored in existing ISP research, for computerized tools that support ISP design. Similarly, [8] Rostami et al., surveyed present studies about the conduct of information security policy (ISP) to scrutinize the amount of proposed manual and computerized support, and also their techniques and procedures. It concluded that for the management of the Information Security Policy generally only manual support is suggested in the prevailing literature. [9] Computerized support is a rarely discussed domain. It proposed for future researchers to further implement computerized tools for the management of Information Security Policy, e.g., procedures that include design science and action research.

It has been found that many small, medium and micro enterprises (SMMEs) do not comply with sound information security governance principles, specifically those principles involved in drafting information security policies and monitoring compliance, mainly as a result of restricted resources and expertise [10]. Research has suggested that this problem occurs worldwide and that the impact it has on SMMEs is great. Another research work is found which introduced a software program to demonstrate the information security governance models practical feasibility, called The Information Security Governance Toolbox (ISGT) [11].

In the studied literature, one of the research artifacts suggested a model to compare the low-level security policy to a high-level security policy on the basis of compliance between them. Another very similar study is found which the same problem is discussed but the solution is very different [12]. The administrative and security metadata was considered while building this framework. The refinement of high-level concepts to was reinforced with the results of refinement calculus so that the refinement patterns and their properties prove to be effective and authentic. The two security policies are said to be in compliance if a valid refinement path can be detected from the high-level security policy to the low-level security policy. This framework could spot defilements of security policies, negligence to complete requirements, and competence and modal conflicts.

Automated systems always prove to be more efficient as compared to manually doing the same task. A study about Distributed security policy conformance - [13], found that manual attempts to audit distributed systems are tedious, error prone, and potentially vulnerable to insider attacks or credential theft. Therefore, it suggested that the formalization of security policies and the use of hardened automated systems that validate compliance can improve the quality and efficiency of this auditing process.

A Systematic Literature Review about Information Security Policy Compliance [13] found that there is a lack of study about an evaluation of information security policy compliance using specific metric and need to enhance the model of information security policy compliance with organizational theories. It suggested for future work to develop instruments that can be used to measure compliance with information security policies.

A comparative study of ontologies-based ISO 27000 series security standards [14] presented security guidelines and best practices in term of concepts and their relationships for effective exploitation, reuse and comprehension of security standards in any organization. It stated that there is still a need to develop a unified security ontology covering all relevant security concepts, incorporating several requirements from ISO 27000 series, following a well-defined methodology and ensuring the assessment and validation of the security ontology. Standards contain of a vast quantity of material. For instance, the international security standard ISO 27000-series comprises of 450 objects with 9 areas of emphasis. Small- and medium sized businesses hardly ever completely apply these security standards which results in a lag to ad-hoc applications [15]. There is no straightforward or simple tool available to be used by small- and medium sized organizations. Impending implementation of industrious tool or technique to enumerate the level of information security is looked-for and along with these, procedures to combine them on the basis of vital security pointers.

Another method which is frequently discussed in the literature studies is the use of benchmarking for selecting a standard for information security policy documents. An an easy approach for organizations to select a suitable information security policy for them is the use of benchmarking [16, 17]. But choosing an appropriate organization as a benchmark is a difficult task because of the dearth of quantifiable procedures for benchmarking. It suggested that scholars should shed light on the subtleties of heterogeneous organizations that share comparable features of ISSP. Another such study proposed an artifact for the benchmarking technique of information security policy. The proposed model enables the execution of effective information security policies. It can be used by the organizations to evaluate and benchmark information security policies [18]. This artifact is abstract as it can be implemented for any security policy within the ISO set of security standards. The artifact can also be applied on different referent groups. Security compliance generally indicates the compliance with industry accepted security standards such as NIST, ISO 270001/27002, HIPAA, PCI, etc. A thesis [19]. proposed a model that measures security compliance of CSP with the major international standard organization against data breaches threat. Semantic similarity measure is used to measure compliance.

2.3.2 Information Security Policy Documents

Writing style and way of communication of the policy to user is also as important as the technical details [20]. A huge number of research artifacts have discussed the important steps in the formulation of an effective security policy and implementation of a security policy document. The important points that must be included in an information security policy document according to the international security standards are discussed in many already existing research discussions and surveys [21–23]. Research discussed essentials of a security policy, its writing procedures and guidelines and its implementation at every level in an organization [a2014impact]. A research study [24] proposed a model for the development of an information security policy in modern organizations based on recommended practices from a sample of certified information security professionals. The model provided relevant guidance for practice and theoretical insight for research. The proposed process model represented a generalized framework rather than a specific model for a single company.

For state-of-the-art ISP development, the focus should shift more toward organization-specific information security needs, as the direction of the current research is still lacking contributions that would show how contextual factors could be successfully integrated into ISP development [25]. Studied literature discussed the need to analyze and validate security policies [26]. It proposed a system to analyze security policies based on deductive spreadsheets using role-based access control. In the e-business arena, firms must have information security policy. The typical objectives of security policy and the technical portions of information security amenities i.e., Non-Reputation, Integrity of Data, Authorization and Privacy Authentication are found in previous research [27]. In addition to that, the technologies to implement these well-known services, I.e., Symmetric cryptography and Asymmetric cryptography, are also discussed.

Enforcing database security policies ensures compliance with regulations that may be governing an organization [28]. This research discussed many solutions for preventing data breaches, one of those solutions is by enforcing database security plans and policies. To ensure routine checks are performed to uncover any deviations from a documented system baseline such as in a System Security Plan (SSP) are reviewed and justified.

Tuyikeze et al., suggested [29] that many organizations are able to define and meet their basic requirements by following a set of reasonable, standard principles in a structured way. Implementing an optimized information security policy is not an easy task, organizations go through some common pitfalls. Following a roadmap for information security policy development might promote sustainability [30]. An Information Security Policy Development Life Cycle (ISP-DLC) was one such proposition. Using this security policy life cycle will provide a framework to help organizations ensure that the necessary steps for security policy development are performed consistently over the life of the policy and that the policies are complied with [31]. An effective information security policy can be very beneficial as it can help to avoid insider threats [32]. The proposed model, for the formulation, implementation and enforcement of an information security policy in an organization, in the studied literature provided the different dimensions that a specific organization needs to take into account during the information security policy development and implementation process. It ensured both comprehensive and sustainable information security policies.

The impact of executing and properly implementing policies and procedures can determine success or failure for information security [33]. Reviewed literature described important topics regarding information security policies, i.e., Developing effective policies and procedures, Internal control, risk assessment, risk control, disaster recovery and business continuity [34].

Previous literature identified information security policy as one of the three key success factors of information systems security, while the other two being management support and information security education, training and awareness. It also discussed that these security policies must be developed properly in order to get complete compliance by employees. A research recommended that the implementation of proposed theoretical models is particularly necessary [35].

Many research articles suggested a generalized policy for information security to be used in organizations. One such research identified potential security policies that can be implemented for cyberspace by organizations. These identified policies can be used as a blueprint for organization cyber security practices [36]. Another research proposed a step-by-step comprehensive process for security policy development and implementation. It discussed the importance of security policy in higher education and how its development different from security policy development of corporate organizations [5].

A research paper explained in detail the different steps in the information security policy development. It proposed a policy design framework for network security [37]. Different security policy development techniques and lifecycles have been compared and exhaustively reviewed in the literature.

2.3.3 Similarity Measurement Techniques

The studied literature also contains research works about the techniques that will be used in this research study. There are primarily four types of tests that can be used to determine document similarity [38]. These are binary similarity models, count similarity models (Jaccard and Cosine measures included), LSA similarity models, ontology-based similarity models. Another survey discussed the use of different text similarity approaches, i.e., String-based, Corpus-based and Knowledge-based [39]. Many metrics, such as Euclidean distance-based metric, Cosine, Jaccard, Dice, JensenShannon Divergence-based metric, have been suggested in recent years to deal with various forms of information retrieval and problems with natural language processing [40], [41]. Among the existing metrics, Cosine, which measures the angle between two vectors, is the most popular one. It is effectively calculated as dot-product of two normalized vectors. [42]. Similarly, cosine similarity, and a mixture of Jaccard similarity and cosine similarity were used in the Jaccard similarity process.

The significance of the similarity of the two names is predicted to increase by integrating the two similarities [43]. Traditional document clustering techniques rely heavily on the presence of keywords and the number of times they appear. The majority of term frequency dependent clustering techniques treat documents as if they were a bag of terms, ignoring the essential relationships between the words in the text. Phrase based clustering techniques also capture only the order in which the words occur in a sentence rather than the semantics behind the words [44, 45]. One more such survey discussed several algorithms of different text similarity approaches, i.e., String-based, Corpus-based and Knowledge-based, including Cosine similarity, Euclidean distance and Jaccard similarity. One more important thing mentioned in this survey is that hybrid text similarity approaches give better results as compared to their results when used separately [46]. There are two types of similarities, one of them is textual and the other one is semantic. Most of these surveys focus on the textual similarities. Existing studies only consider the textual similarity but do not consider the semantics behind the data [47]. Some researchers even suggested new and improved versions of the already present techniques. For instance, a new method for calculating semantic similarities between documents was proposed. It was based on cosine similarity calculation between concept vectors of documents obtained from a taxonomy of words that captures IS-A relations [48]. It had same time complexity as cosine similarity but gave better results. In another research, a new similarity measurement technique, called improved sqrt-cosine (ISC) similarity, which was based on Hellinger distance, was proposed [49]. It was very similar to cosine similarity approach but instead of using Euclidean distance it used Hellinger distance and performed very well for high dimensional data. Another study analyzed several different similarity measures by applying them on different kinds of datasets and concluded that there were no or very less noise points in clusters created by Jaccard and Cosine functions, Euclidean function had some noise points while clusters built using correlation functions had a lot of noise points [50].

2.4 Conclusions

After the exhaustive research, it can be concluded that the problem discussed in this study is valid. The answers of survey questions found in the above literature review can be concluded as:

1. The very few available solutions are either too complex and expensive or not very abstract. No concrete study has been found that suggests any technique to find compliance of information security policy documents to a standard template.

2. Even though technique for compliance identification of ISPD is not found in literature but compliance identification between documents is a commonly discussed topic and some common Similarity Measures are being used in them. [5] checks compliance of a document with a template but the document being checked is a business contract. [12] checks compliance of a low-level security policy document to a high-level security policy document

by the use of calculus. Literature also contains research that discusses several techniques.

3. There are many artifacts containing comparisons of different policy development lifecycles and several surveys on key points that must be in an organizations security policy document according to the international security standards.

4. Many studies which suggest developing a computerized tool for information security policy documents. Therefore, this research problem proves to be an important addition to the field of information technology.

Chapter 3

Proposed Research Technique

3.1 Introduction

This section proposes a technique to solve the problem of non-compliance of information security policy documents with the necessary security standard by calculating the similarity between a standard document with various test documents. Standard document is a document that is being considered as a reference to find the similarity between two documents. Test document is the document which will be compared to reference document to obtain its similarity score with the standard document. The documents considered for this study are chosen from the health domain. The comparison is done both rhetorically and on the basis of contents, to find out the extent of similarity score.

3.2 Experimental Setup

3.2.1 Programming Language

For the implementation of this experiment, Python programing language is used. It is a powerful language as it contains several built-in libraries for the purpose of text manipulation and comparison. It is most suitable because along with the useful libraries it is also very easy to implement and run with minimum system requirements.

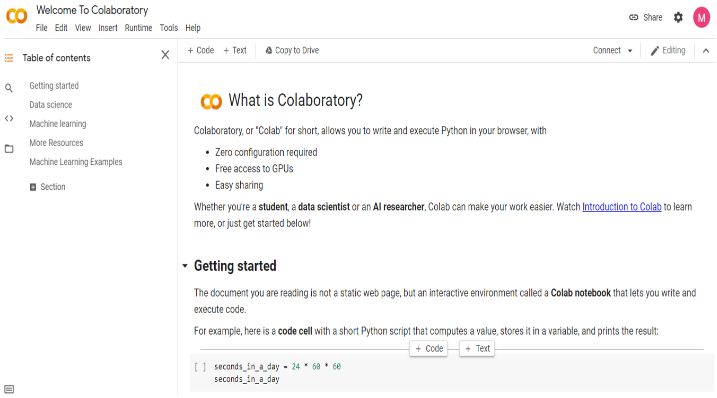
3.2.1.1 Libraries and Methods

The Natural Language Toolkit [1] commonly abbreviated as nltk is a package which contains many methods and libraries for statistical and lexical analysis of natural language. The implicit methods for preprocessing of text, i.e., removal of stop words, tokenization (nltk.tokenize), dictionary for synonym identification (wordnet) etc., come within its installation package. Wordnet is a lexical dictionary of English language. It groups different words into synsets which are similar to each other. [2]

The python library python-docx [3] and the PDF Toolkit abbreviated as PyPDF2 [4] consist of various useful functionalities like extracting distinguished information from documents, splitting them in parts, merging or cropping them etc. It is used to extract headings and text from the documents.

3.2.2 Tools Used

The tool used for the implementation of this experiment is Google Colab.



What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ]  seconds_in_a_day = 24 * 60 * 60
     seconds_in_a_day
```

FIGURE 3.1: Frontend of Google Colab

[1] https://www.nltk.org/

[2] https://www.nltk.org/howto/wordnet.html

[3] https://python-docx.readthedocs.io/en/latest/

[4] https://pythonhosted.org/PyPDF2/

It is one of the major useful platforms for implementing Python projects as it has pre-installed libraries and it allows users to save the code in the form of Jupyter notebooks on cloud and access or execute it anywhere through a browser.

3.2.3 Machine Configuration

The system used to execute the project is Dell i5
Processor: 5th Generation Intel Core i5-5200U Processor (3M Cache, up to 2.70 GHz)
Operating System: Windows 10 Pro
Web browser: Google Chrome

3.3 Selection of Documents

All of the documents involved in this experiment come from the domain of health sector. Information Security Policy Documents from various reputable and government organizations are easily available on the internet. The importance of security standards in health domain is underrated as a single security breach in such a system can put lives of stake. Focusing on a single domain is beneficial as it is giving more accurate results. 5 such documents are selected for this experiment. One of them is chosen as the standard template, rest of them are kept for the testing. All of the documents are docx files.

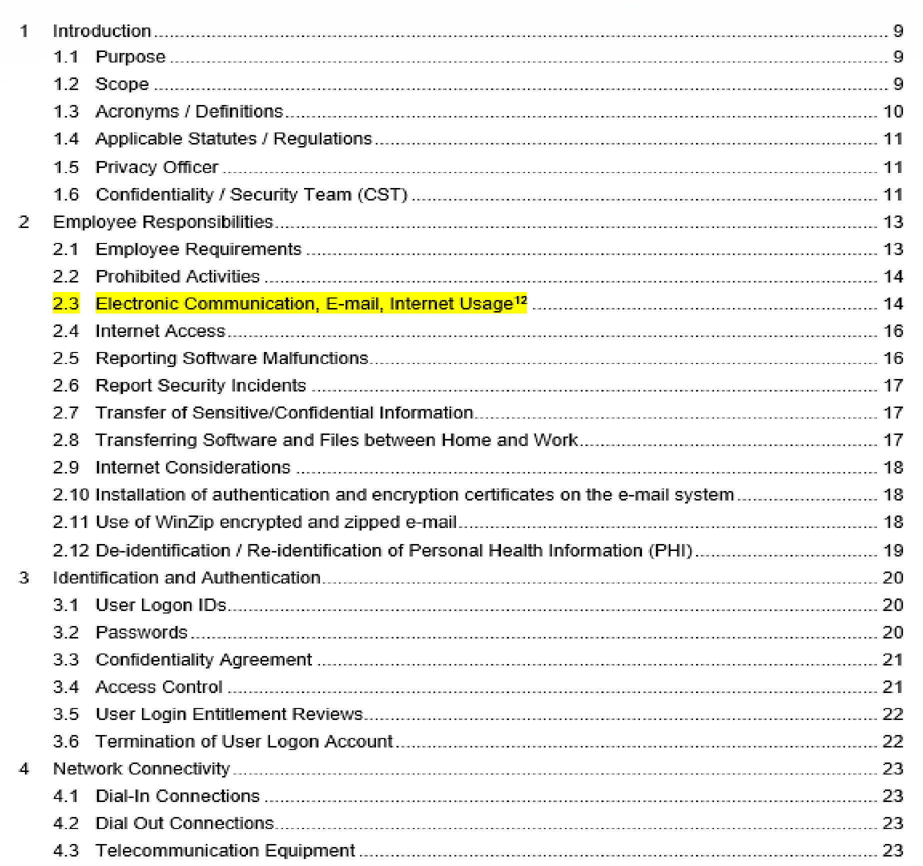
3.3.1 Standard Template Document[5]

The document selected as standard is a template provided by National Learning Consortium (NLC) [6] and it is developed by the Privacy & Security team of Health Information Technology Research Center (HITRC). This American organization is known for designing knowledge and resources to support healthcare providers

[5] https://www.healthit.gov/sites/default/files/tools/info_security_policy_template_v1_0.docx
[6] https://www.healthit.gov/topic/health-it-resources

and health IT professionals. The standard template was thoroughly reviewed for any shortcomings on the basis of reviewed literature [24, 51] about important security standards in policy making and available knowledge about international standards of security policies. There are various reasons of choosing this document as standard, the first and foremost being the length of this document. As it is the most comprehensive, it contains maximum of the important headings that must be present in any information security document of a healthcare organization. This document is found close to the knowledge gathered about international security standards in this study. The organization that created the document is a reputable government organization known for producing documents and tools that aid in increasing the security of healthcare organizations.

- Last reviewed in 2011
- Total number of pages: 94



1 Introduction 9
1.1 Purpose 9
1.2 Scope 9
1.3 Acronyms / Definitions 10
1.4 Applicable Statutes / Regulations 11
1.5 Privacy Officer 11
1.6 Confidentiality / Security Team (CST) 11
2 Employee Responsibilities 13
2.1 Employee Requirements 13
2.2 Prohibited Activities 14
2.3 Electronic Communication, E-mail, Internet Usage[12] 14
2.4 Internet Access 16
2.5 Reporting Software Malfunctions 16
2.6 Report Security Incidents 17
2.7 Transfer of Sensitive/Confidential Information 17
2.8 Transferring Software and Files between Home and Work 17
2.9 Internet Considerations 18
2.10 Installation of authentication and encryption certificates on the e-mail system 18
2.11 Use of WinZip encrypted and zipped e-mail 18
2.12 De-identification / Re-identification of Personal Health Information (PHI) 19
3 Identification and Authentication 20
3.1 User Logon IDs 20
3.2 Passwords 20
3.3 Confidentiality Agreement 21
3.4 Access Control 21
3.5 User Login Entitlement Reviews 22
3.6 Termination of User Logon Account 22
4 Network Connectivity 23
4.1 Dial-In Connections 23
4.2 Dial Out Connections 23
4.3 Telecommunication Equipment 23

Figure 3.2: Table of Contents (1) of Standard Document

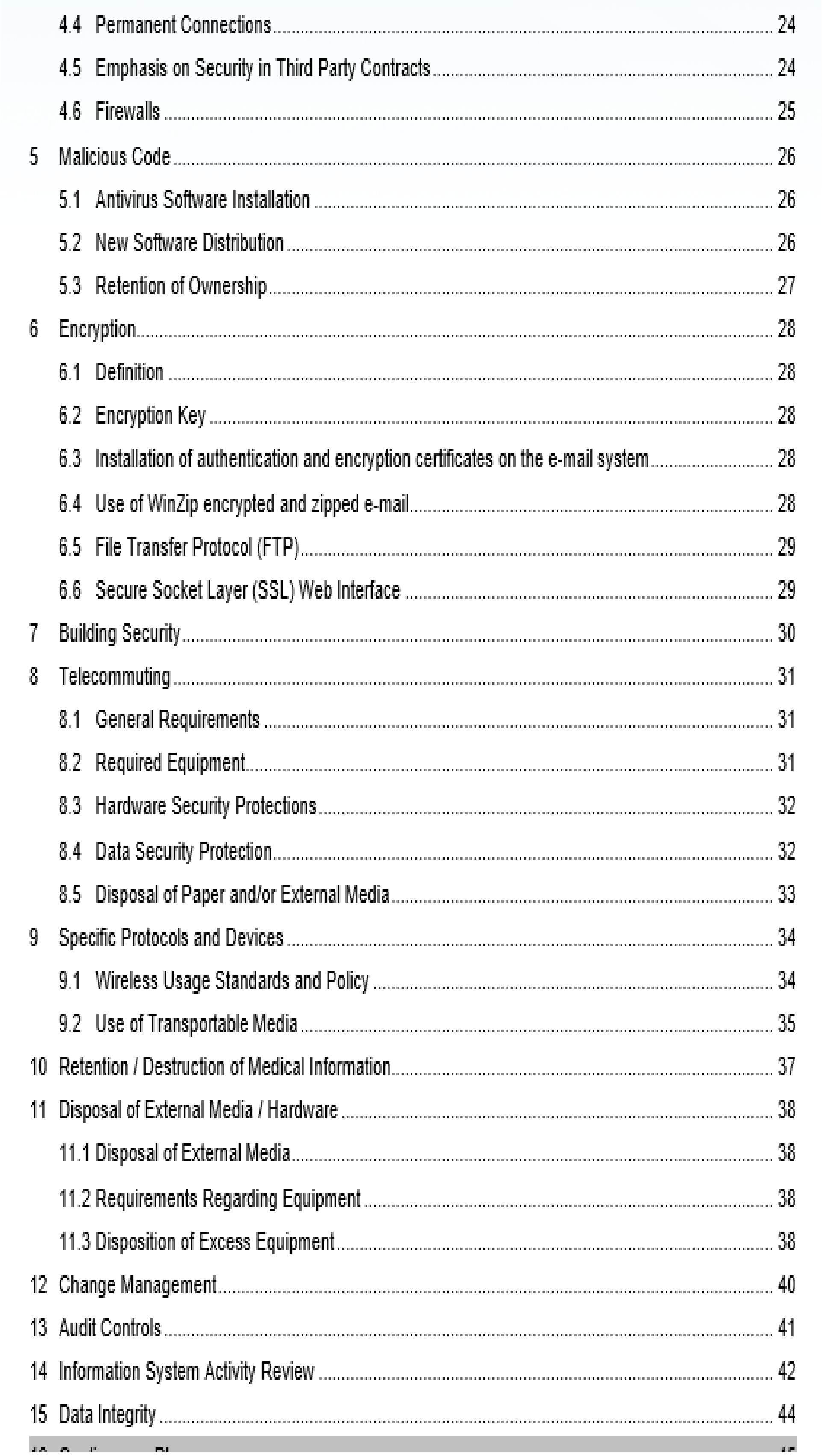
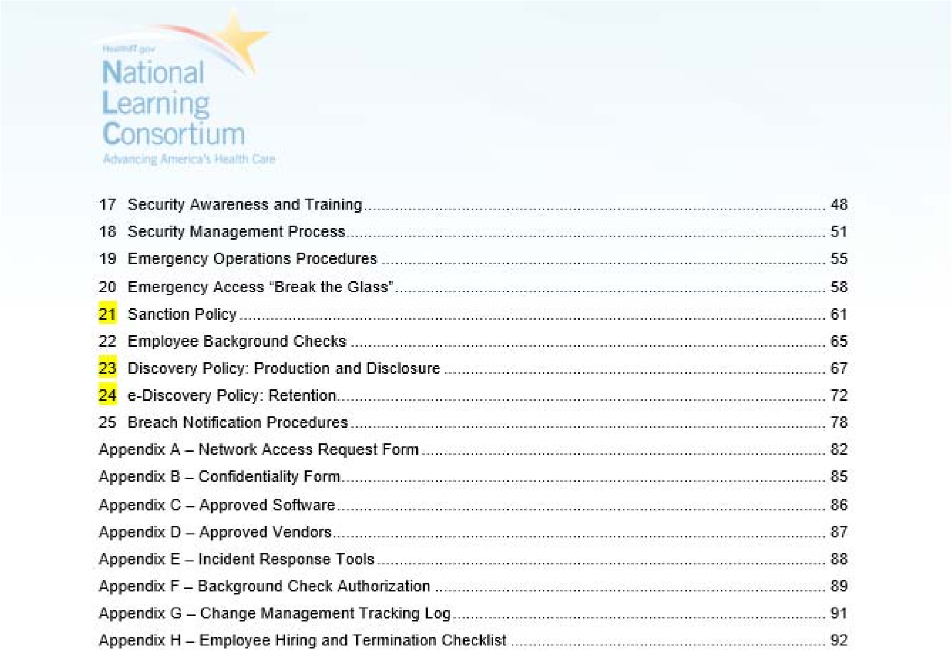
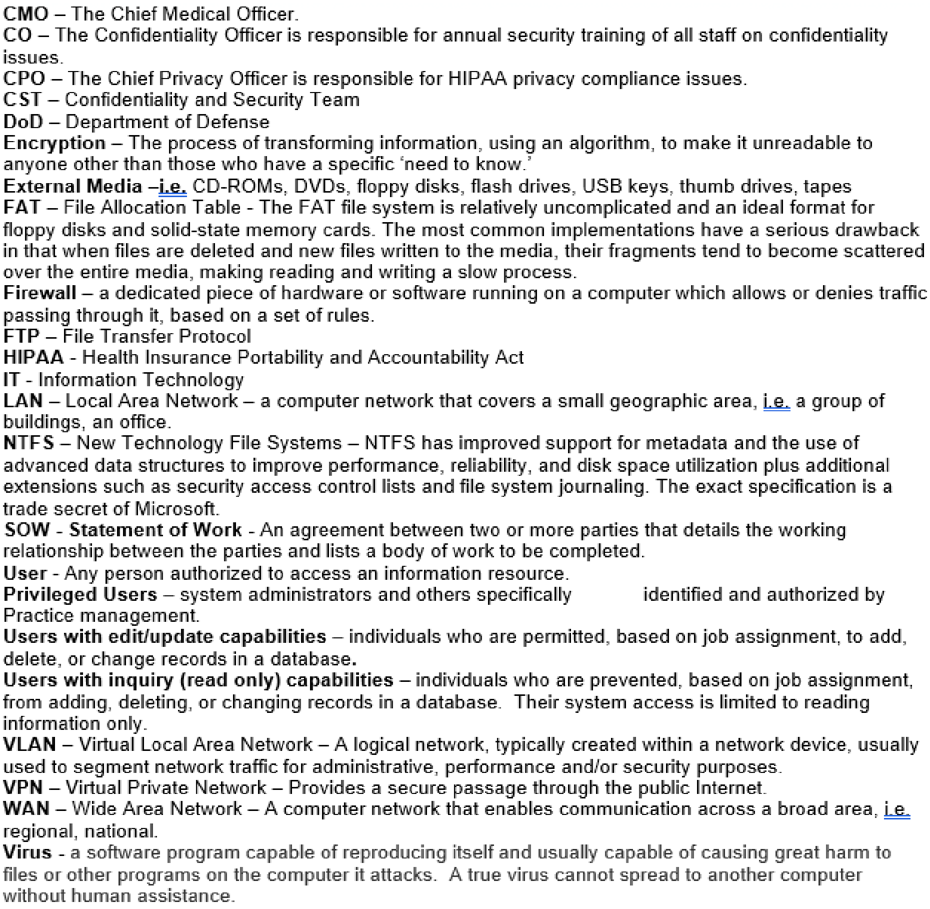


Figure 3.3: Table of Contents (2) of Standard Document



HealthIT.gov
National Learning Consortium
Advancing America's Health Care

Figure 3.4: Table of Contents (3) of Standard Document



CMO – The Chief Medical Officer.
CO – The Confidentiality Officer is responsible for annual security training of all staff on confidentiality issues.
CPO – The Chief Privacy Officer is responsible for HIPAA privacy compliance issues.
CST – Confidentiality and Security Team
DoD – Department of Defense
Encryption – The process of transforming information, using an algorithm, to make it unreadable to anyone other than those who have a specific ‘need to know.’
External Media –i.e. CD-ROMs, DVDs, floppy disks, flash drives, USB keys, thumb drives, tapes
FAT – File Allocation Table - The FAT file system is relatively uncomplicated and an ideal format for floppy disks and solid-state memory cards. The most common implementations have a serious drawback in that when files are deleted and new files written to the media, their fragments tend to become scattered over the entire media, making reading and writing a slow process.
Firewall – a dedicated piece of hardware or software running on a computer which allows or denies traffic passing through it, based on a set of rules.
FTP – File Transfer Protocol
HIPAA - Health Insurance Portability and Accountability Act
IT - Information Technology
LAN – Local Area Network – a computer network that covers a small geographic area, i.e. a group of buildings, an office.
NTFS – New Technology File Systems – NTFS has improved support for metadata and the use of advanced data structures to improve performance, reliability, and disk space utilization plus additional extensions such as security access control lists and file system journaling. The exact specification is a trade secret of Microsoft.
SOW - **Statement of Work** - An agreement between two or more parties that details the working relationship between the parties and lists a body of work to be completed.
User - Any person authorized to access an information resource.
Privileged Users – system administrators and others specifically identified and authorized by Practice management.
Users with edit/update capabilities – individuals who are permitted, based on job assignment, to add, delete, or change records in a database.
Users with inquiry (read only) capabilities – individuals who are prevented, based on job assignment, from adding, deleting, or changing records in a database. Their system access is limited to reading information only.
VLAN – Virtual Local Area Network – A logical network, typically created within a network device, usually used to segment network traffic for administrative, performance and/or security purposes.
VPN – Virtual Private Network – Provides a secure passage through the public Internet.
WAN – Wide Area Network – A computer network that enables communication across a broad area, i.e. regional, national.
Virus - a software program capable of reproducing itself and usually capable of causing great harm to files or other programs on the computer it attacks. A true virus cannot spread to another computer without human assistance.

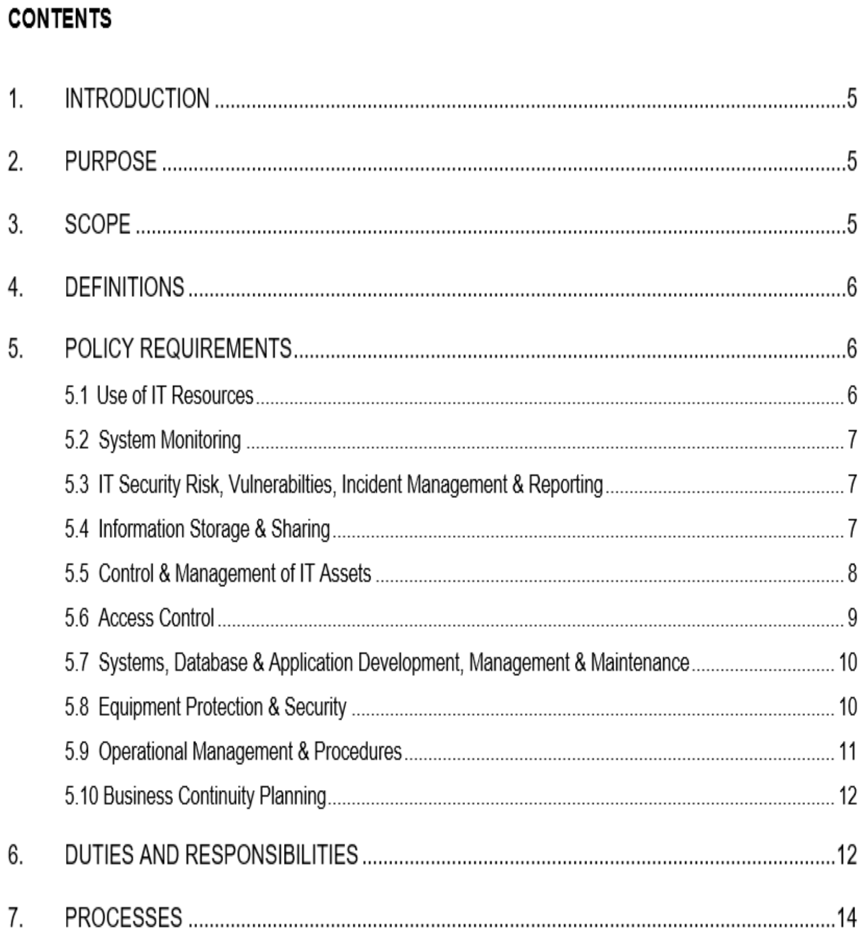
Figure 3.5: Table of Contents (4) of Standard Document

3.3.2 Test Documents

The details of the 10 test documents are as follows:

Test 1 [7]**:** IT security policy document of Portsmouth Hospital NHS Scotland Information Security Policy document. [8]

- Based on international security standard ISO17799
- Last reviewed in 2020
- Total number of pages: 23



CONTENTS

1. INTRODUCTION 5
2. PURPOSE 5
3. SCOPE 5
4. DEFINITIONS 6
5. POLICY REQUIREMENTS 6
   5.1 Use of IT Resources 6
   5.2 System Monitoring 7
   5.3 IT Security Risk, Vulnerabilties, Incident Management & Reporting 7
   5.4 Information Storage & Sharing 7
   5.5 Control & Management of IT Assets 8
   5.6 Access Control 9
   5.7 Systems, Database & Application Development, Management & Maintenance 10
   5.8 Equipment Protection & Security 10
   5.9 Operational Management & Procedures 11
   5.10 Business Continuity Planning 12
6. DUTIES AND RESPONSIBILITIES 12
7. PROCESSES 14

FIGURE 3.6: Table of Contents (1) of Test1 Document

[7] https://www.porthosp.nhs.uk/about-us/policies-and-guidelines/policies/Management/IT%20Security%20Policy.docx

[8] https://www.porthosp.nhs.uk/

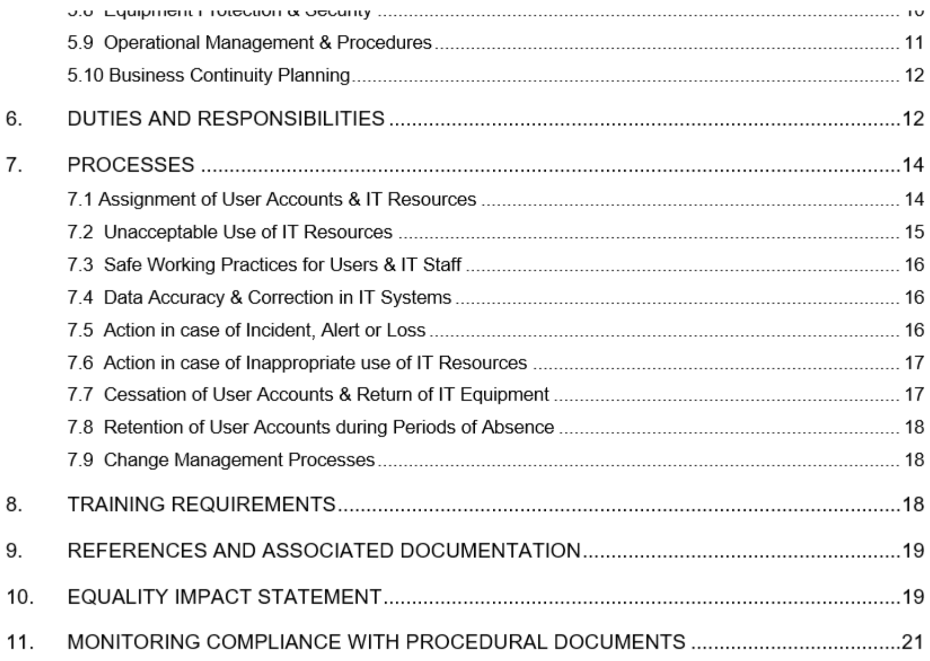
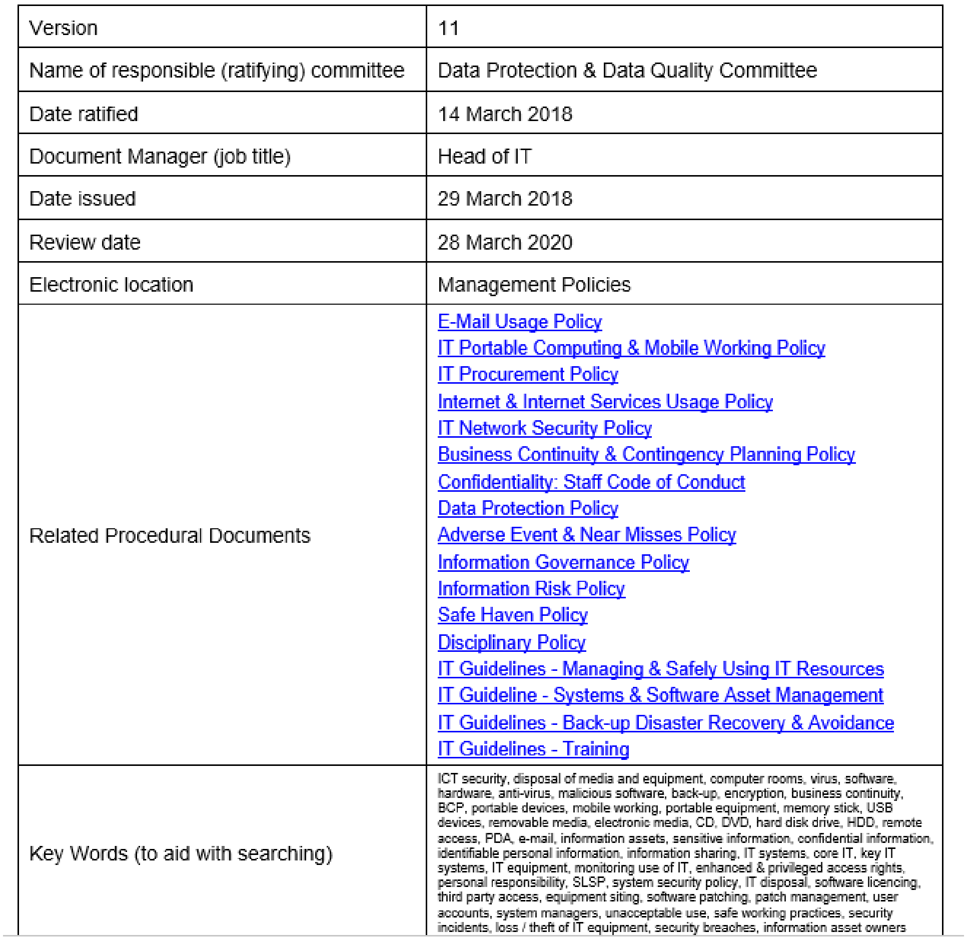


FIGURE 3.7: Table of Contents (2) of Test1 Document

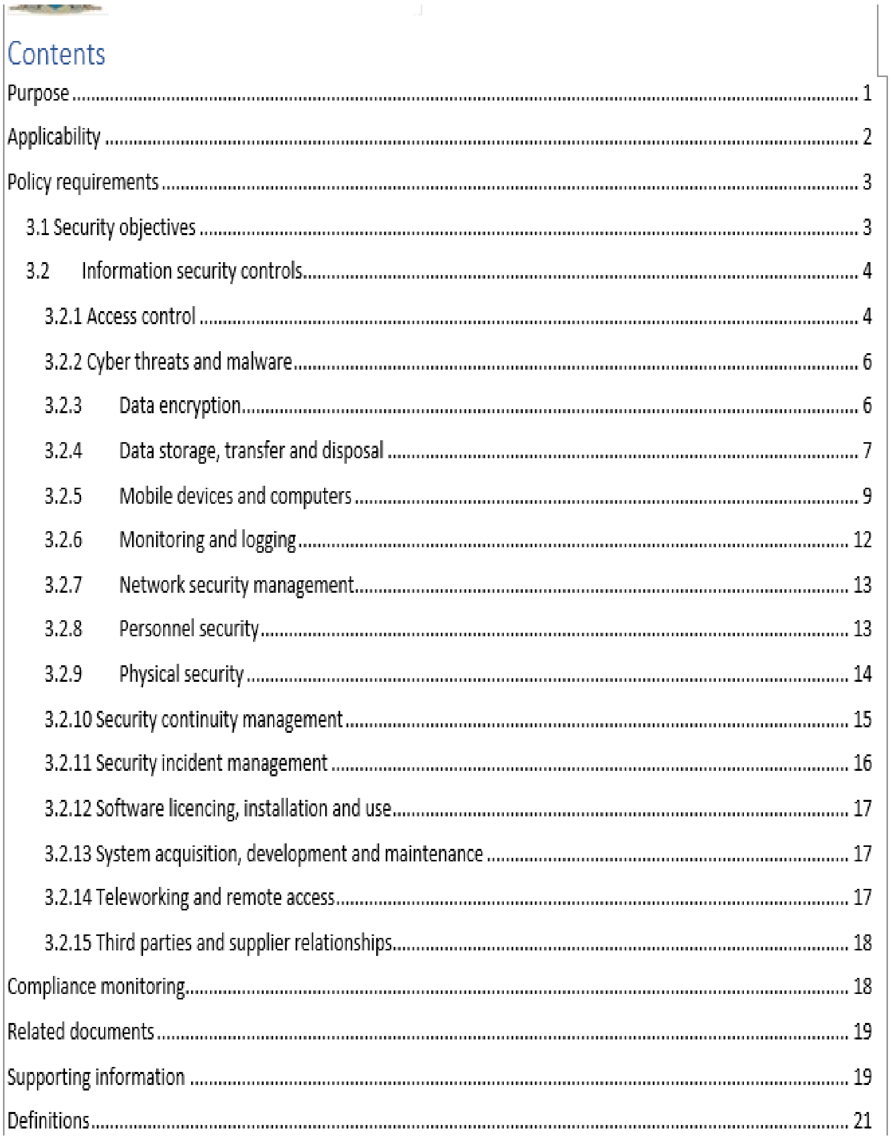


Version	11
Name of responsible (ratifying) committee	Data Protection & Data Quality Committee
Date ratified	14 March 2018
Document Manager (job title)	Head of IT
Date issued	29 March 2018
Review date	28 March 2020
Electronic location	Management Policies
Related Procedural Documents	E-Mail Usage Policy IT Portable Computing & Mobile Working Policy IT Procurement Policy Internet & Internet Services Usage Policy IT Network Security Policy Business Continuity & Contingency Planning Policy Confidentiality: Staff Code of Conduct Data Protection Policy Adverse Event & Near Misses Policy Information Governance Policy Information Risk Policy Safe Haven Policy Disciplinary Policy IT Guidelines - Managing & Safely Using IT Resources IT Guideline - Systems & Software Asset Management IT Guidelines - Back-up Disaster Recovery & Avoidance IT Guidelines - Training
Key Words (to aid with searching)	ICT security, disposal of media and equipment, computer rooms, virus, software, hardware, anti-virus, malicious software, back-up, encryption, business continuity, BCP, portable devices, mobile working, portable equipment, memory stick, USB devices, removable media, electronic media, CD, DVD, hard disk drive, HDD, remote access, PDA, e-mail, information assets, sensitive information, confidential information, identifiable personal information, information sharing, IT systems, core IT, key IT systems, IT equipment, monitoring use of IT, enhanced & privileged access rights, personal responsibility, SLSP, system security policy, IT disposal, software licencing, third party access, equipment siting, software patching, patch management, user accounts, system managers, unacceptable use, safe working practices, security incidents, loss / theft of IT equipment, security breaches, information asset owners

FIGURE 3.8: Table of Contents (2) of Test1 Document

Test 2 [9]**:** IT security policy document of department of health of Government of Western Australia. [10]

- Based on international security standard ISO/IEC 27002
- Last reviewed in 2020
- Total number of pages: 24



Contents

Purpose 1

Applicability 2

Policy requirements 3

3.1 Security objectives 3

3.2 Information security controls 4

3.2.1 Access control 4

3.2.2 Cyber threats and malware 6

3.2.3 Data encryption 6

3.2.4 Data storage, transfer and disposal 7

3.2.5 Mobile devices and computers 9

3.2.6 Monitoring and logging 12

3.2.7 Network security management 13

3.2.8 Personnel security 13

3.2.9 Physical security 14

3.2.10 Security continuity management 15

3.2.11 Security incident management 16

3.2.12 Software licencing, installation and use 17

3.2.13 System acquisition, development and maintenance 17

3.2.14 Teleworking and remote access 17

3.2.15 Third parties and supplier relationships 18

Compliance monitoring 18

Related documents 19

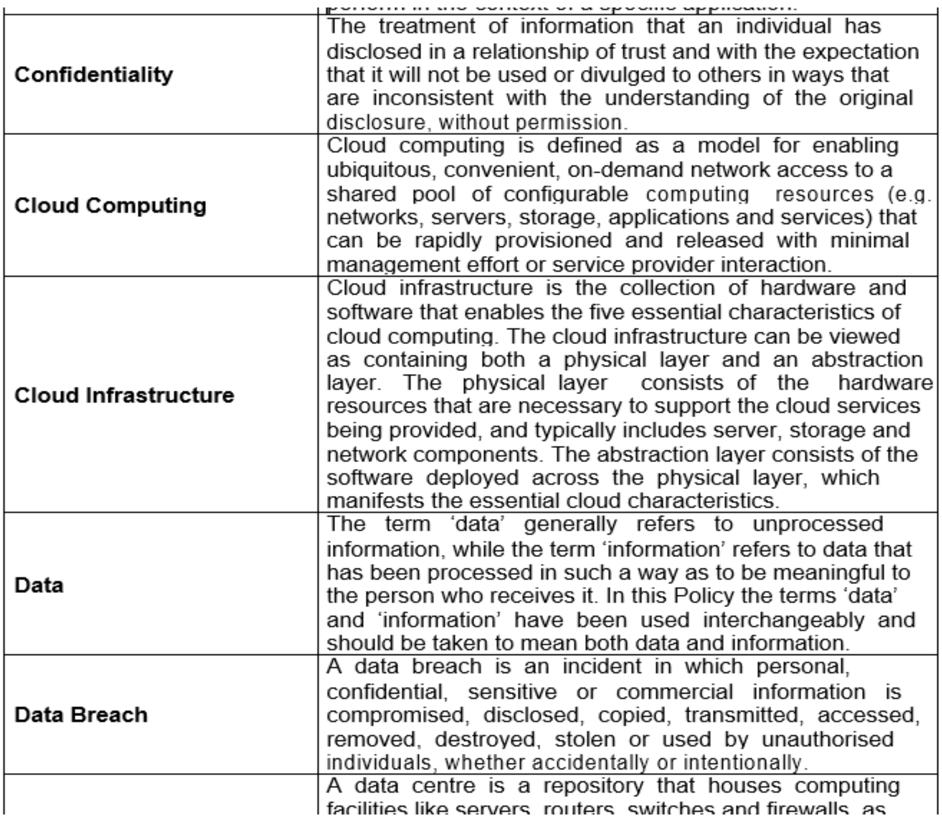
Supporting information 19

Definitions 21

Figure 3.9: Table of Contents (1) of Test2 Document

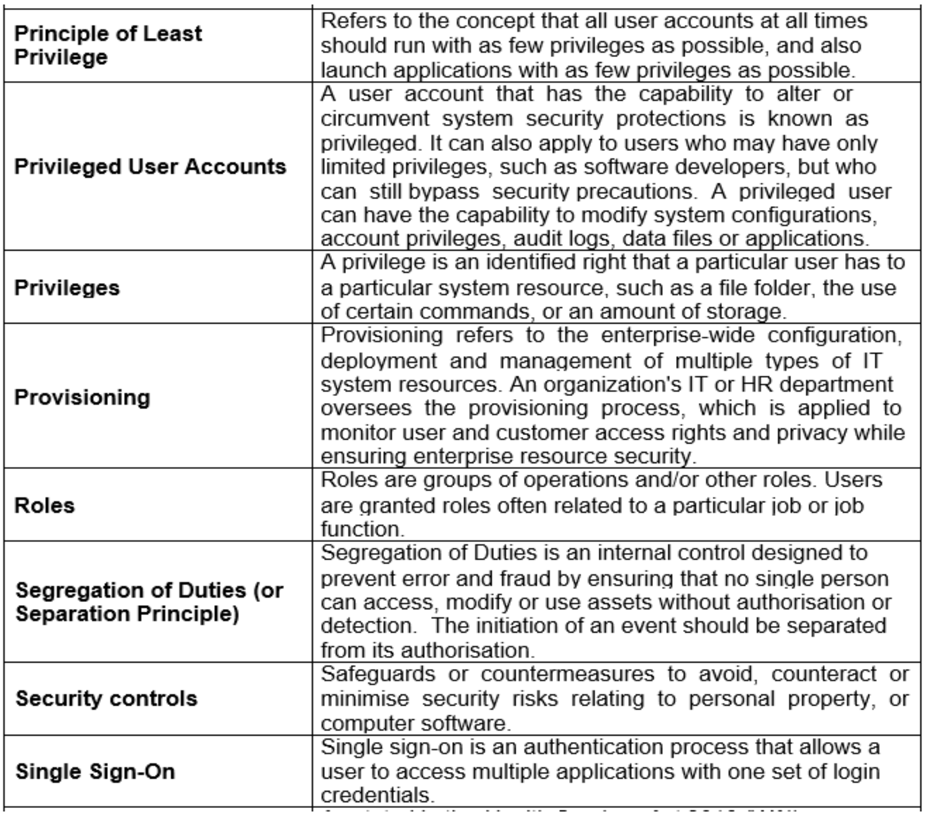
[9] https://ww2.health.wa.gov.au/About-us/Policy-frameworks/Information-Management

[10] https://ww2.health.wa.gov.au/



	perform in the context of a specific application.
Confidentiality	The treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be used or divulged to others in ways that are inconsistent with the understanding of the original disclosure, without permission.
Cloud Computing	Cloud computing is defined as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
Cloud Infrastructure	Cloud infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing. The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources that are necessary to support the cloud services being provided, and typically includes server, storage and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics.
Data	The term 'data' generally refers to unprocessed information, while the term 'information' refers to data that has been processed in such a way as to be meaningful to the person who receives it. In this Policy the terms 'data' and 'information' have been used interchangeably and should be taken to mean both data and information.
Data Breach	A data breach is an incident in which personal, confidential, sensitive or commercial information is compromised, disclosed, copied, transmitted, accessed, removed, destroyed, stolen or used by unauthorised individuals, whether accidentally or intentionally.
	A data centre is a repository that houses computing facilities like servers, routers, switches and firewalls, as

Figure 3.10: Table of Contents (2) of Test2 Document

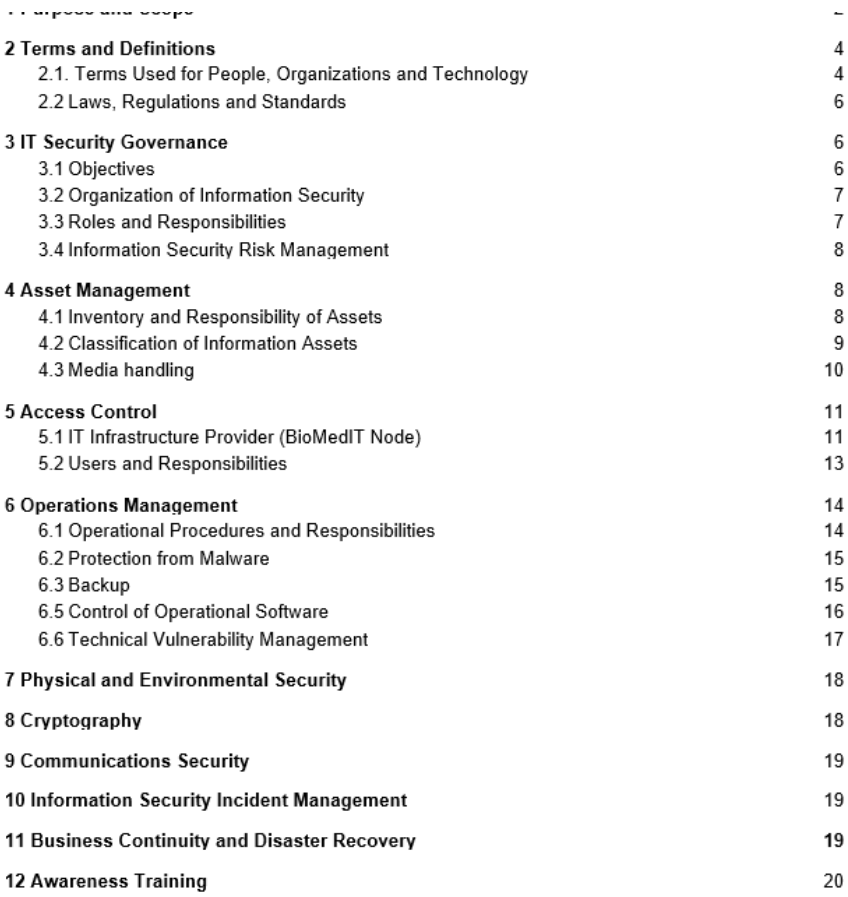


Principle of Least Privilege	Refers to the concept that all user accounts at all times should run with as few privileges as possible, and also launch applications with as few privileges as possible.
Privileged User Accounts	A user account that has the capability to alter or circumvent system security protections is known as privileged. It can also apply to users who may have only limited privileges, such as software developers, but who can still bypass security precautions. A privileged user can have the capability to modify system configurations, account privileges, audit logs, data files or applications.
Privileges	A privilege is an identified right that a particular user has to a particular system resource, such as a file folder, the use of certain commands, or an amount of storage.
Provisioning	Provisioning refers to the enterprise-wide configuration, deployment and management of multiple types of IT system resources. An organization's IT or HR department oversees the provisioning process, which is applied to monitor user and customer access rights and privacy while ensuring enterprise resource security.
Roles	Roles are groups of operations and/or other roles. Users are granted roles often related to a particular job or job function.
Segregation of Duties (or Separation Principle)	Segregation of Duties is an internal control designed to prevent error and fraud by ensuring that no single person can access, modify or use assets without authorisation or detection. The initiation of an event should be separated from its authorisation.
Security controls	Safeguards or countermeasures to avoid, counteract or minimise security risks relating to personal property, or computer software.
Single Sign-On	Single sign-on is an authentication process that allows a user to access multiple applications with one set of login credentials.

Figure 3.11: Table of Contents (3) of Test2 Document

Test 3 [11]**:** IT security policy document of Swiss Personalized Health Network [12]

- Based on SPHN Ethical Framework for Responsible Data Processing
- Last reviewed in 2018
- Total number of pages: 21



1 Purpose and Scope 2

2 Terms and Definitions 4
2.1. Terms Used for People, Organizations and Technology 4
2.2 Laws, Regulations and Standards 6

3 IT Security Governance 6
3.1 Objectives 6
3.2 Organization of Information Security 7
3.3 Roles and Responsibilities 7
3.4 Information Security Risk Management 8

4 Asset Management 8
4.1 Inventory and Responsibility of Assets 8
4.2 Classification of Information Assets 9
4.3 Media handling 10

5 Access Control 11
5.1 IT Infrastructure Provider (BioMedIT Node) 11
5.2 Users and Responsibilities 13

6 Operations Management 14
6.1 Operational Procedures and Responsibilities 14
6.2 Protection from Malware 15
6.3 Backup 15
6.5 Control of Operational Software 16
6.6 Technical Vulnerability Management 17

7 Physical and Environmental Security 18

8 Cryptography 18

9 Communications Security 19

10 Information Security Incident Management 19

11 Business Continuity and Disaster Recovery 19

12 Awareness Training 20

Figure 3.12: Table of Contents of Test3 Document

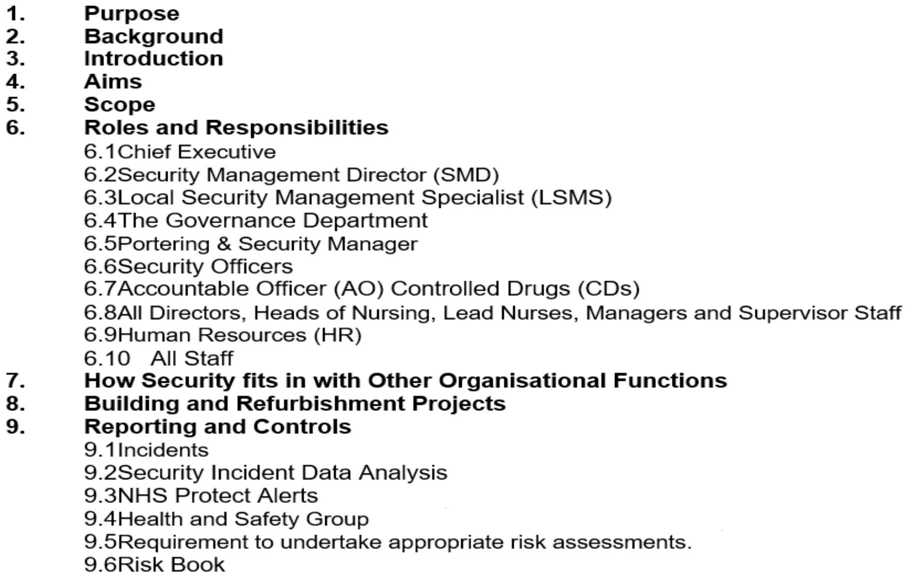
Test 4[13]**:**IT security policy document of Mid Essex Hospital Services NHS Trust [14]

- Based on NHSLA standards 3.9

[11] https://sphn.ch/wp-content/uploads/2020/01/sphn_information_security_policy_v1.pdf
[12] https://sphn.ch/
[13] https://www.scribd.com/document/395625271/BASIC-SECURITY-MEASURES-FOR-GUARDS-IN-FACILITY
[14] https://www.meht.nhs.uk/

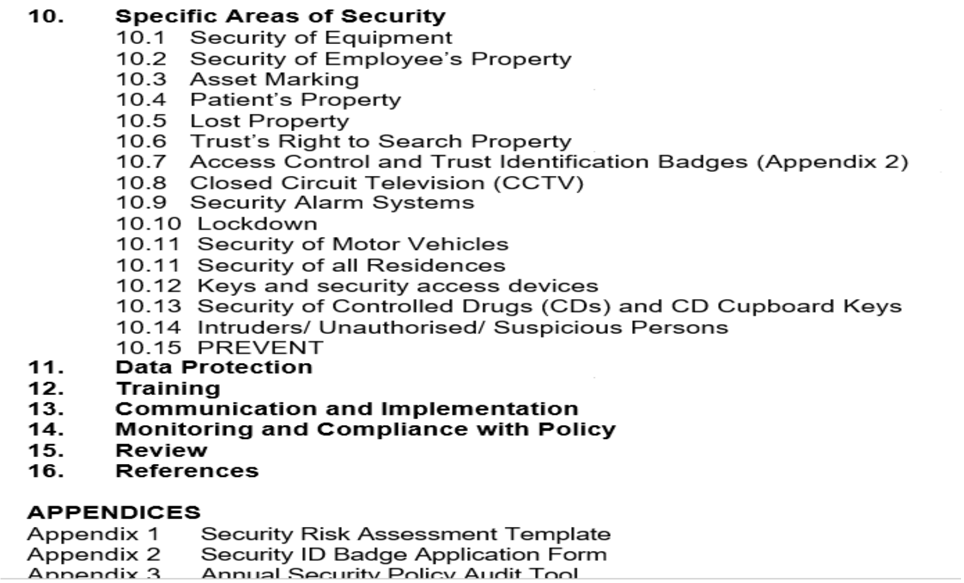
- Last reviewed in 2014

- Total number of pages: 18



1. **Purpose**
2. **Background**
3. **Introduction**
4. **Aims**
5. **Scope**
6. **Roles and Responsibilities**
   6.1Chief Executive
   6.2Security Management Director (SMD)
   6.3Local Security Management Specialist (LSMS)
   6.4The Governance Department
   6.5Portering & Security Manager
   6.6Security Officers
   6.7Accountable Officer (AO) Controlled Drugs (CDs)
   6.8All Directors, Heads of Nursing, Lead Nurses, Managers and Supervisor Staff
   6.9Human Resources (HR)
   6.10 All Staff
7. **How Security fits in with Other Organisational Functions**
8. **Building and Refurbishment Projects**
9. **Reporting and Controls**
   9.1Incidents
   9.2Security Incident Data Analysis
   9.3NHS Protect Alerts
   9.4Health and Safety Group
   9.5Requirement to undertake appropriate risk assessments.
   9.6Risk Book

Figure 3.13: Table of Contents (1) of Test4 Document



10. **Specific Areas of Security**
    10.1 Security of Equipment
    10.2 Security of Employee's Property
    10.3 Asset Marking
    10.4 Patient's Property
    10.5 Lost Property
    10.6 Trust's Right to Search Property
    10.7 Access Control and Trust Identification Badges (Appendix 2)
    10.8 Closed Circuit Television (CCTV)
    10.9 Security Alarm Systems
    10.10 Lockdown
    10.11 Security of Motor Vehicles
    10.11 Security of all Residences
    10.12 Keys and security access devices
    10.13 Security of Controlled Drugs (CDs) and CD Cupboard Keys
    10.14 Intruders/ Unauthorised/ Suspicious Persons
    10.15 PREVENT
11. **Data Protection**
12. **Training**
13. **Communication and Implementation**
14. **Monitoring and Compliance with Policy**
15. **Review**
16. **References**

APPENDICES
Appendix 1 Security Risk Assessment Template
Appendix 2 Security ID Badge Application Form
Appendix 3 Annual Security Policy Audit Tool

Figure 3.14: Table of Contents (2) of Test4 Document

Test 5 [15]**:** IT security policy document of eHealth Ontario, a 21st-century government agency providing high-quality health care services [16]

[15] https://www.ehealthontario.on.ca/files/public/support/Information_Security_Policy_EN.pdf
[16] https://ehealthontario.on.ca/en

- Last reviewed in 2019
- Total number of pages: 22

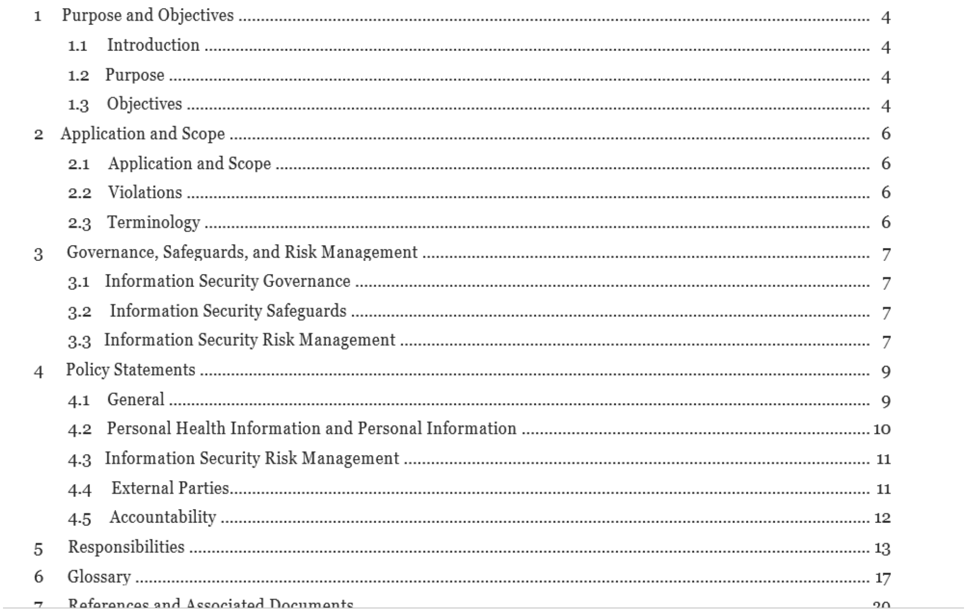
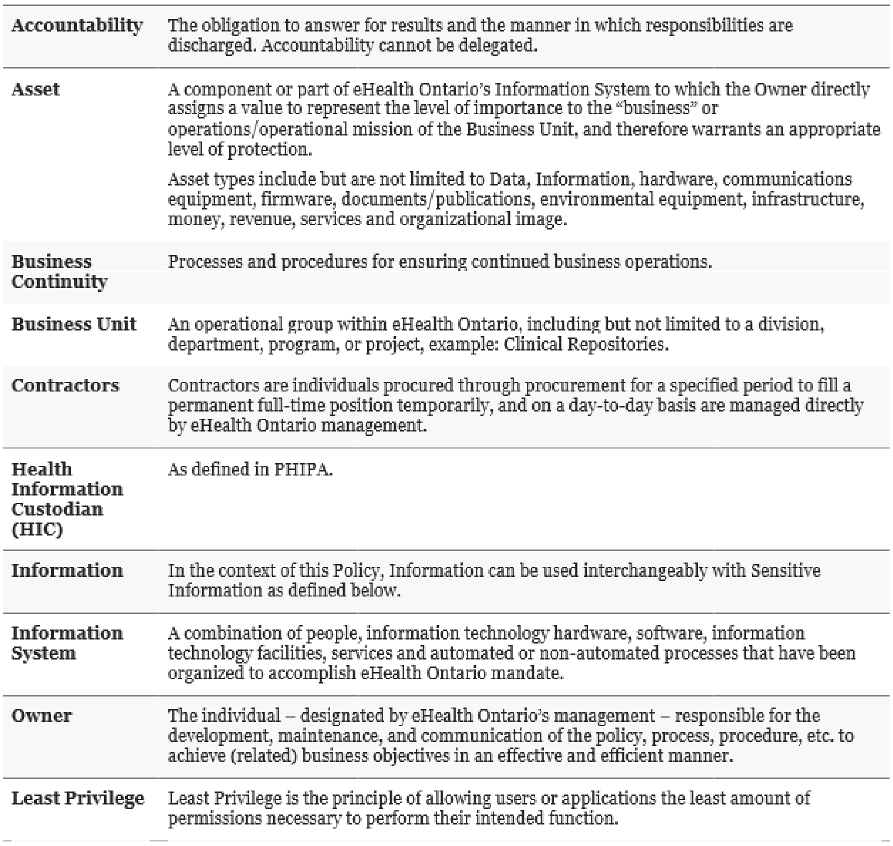
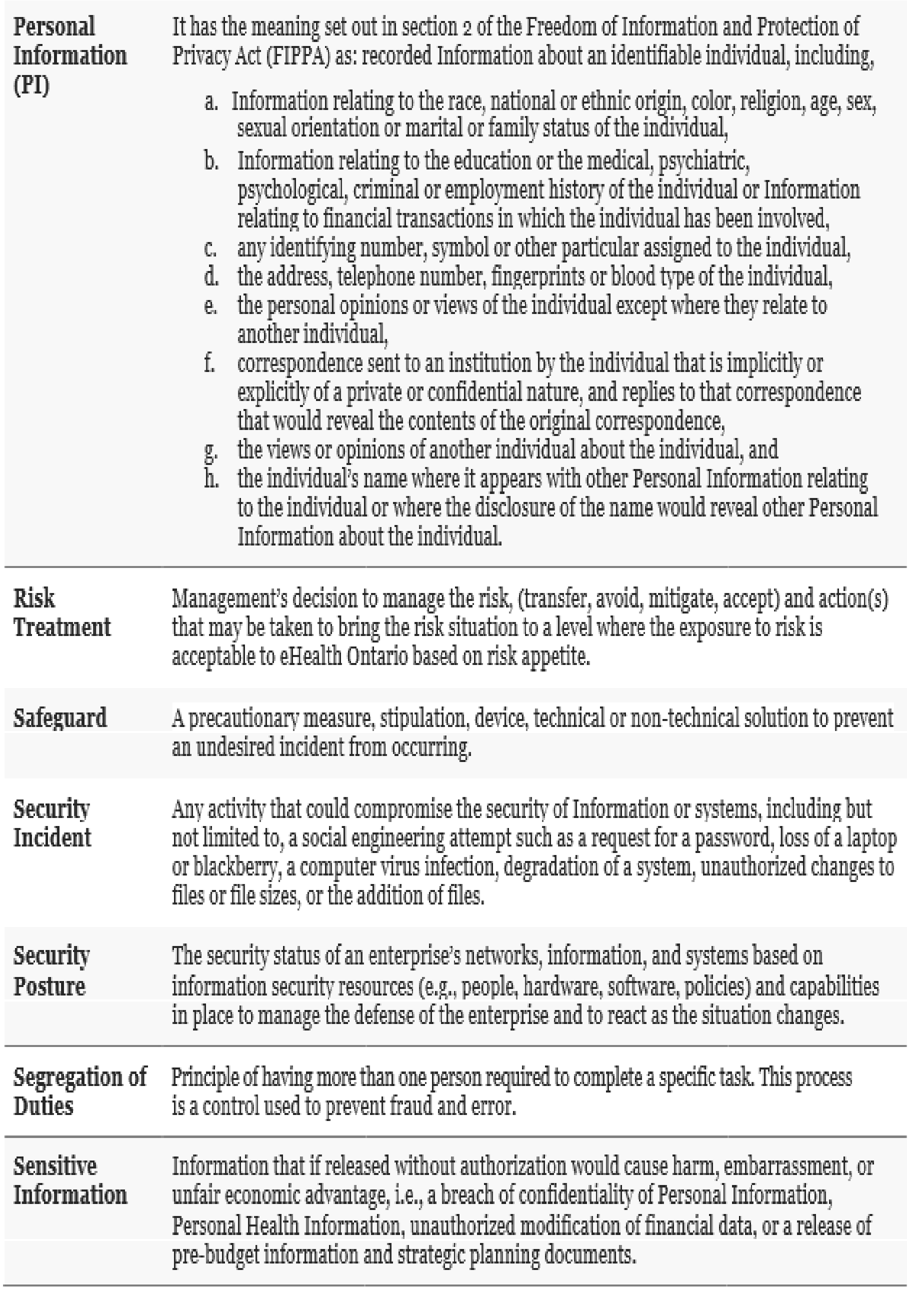


FIGURE 3.15: Table of Contents (1) of Test5 Document



Accountability	The obligation to answer for results and the manner in which responsibilities are discharged. Accountability cannot be delegated.
Asset	A component or part of eHealth Ontario's Information System to which the Owner directly assigns a value to represent the level of importance to the "business" or operations/operational mission of the Business Unit, and therefore warrants an appropriate level of protection. Asset types include but are not limited to Data, Information, hardware, communications equipment, firmware, documents/publications, environmental equipment, infrastructure, money, revenue, services and organizational image.
Business Continuity	Processes and procedures for ensuring continued business operations.
Business Unit	An operational group within eHealth Ontario, including but not limited to a division, department, program, or project, example: Clinical Repositories.
Contractors	Contractors are individuals procured through procurement for a specified period to fill a permanent full-time position temporarily, and on a day-to-day basis are managed directly by eHealth Ontario management.
Health Information Custodian (HIC)	As defined in PHIPA.
Information	In the context of this Policy, Information can be used interchangeably with Sensitive Information as defined below.
Information System	A combination of people, information technology hardware, software, information technology facilities, services and automated or non-automated processes that have been organized to accomplish eHealth Ontario mandate.
Owner	The individual – designated by eHealth Ontario's management – responsible for the development, maintenance, and communication of the policy, process, procedure, etc. to achieve (related) business objectives in an effective and efficient manner.
Least Privilege	Least Privilege is the principle of allowing users or applications the least amount of permissions necessary to perform their intended function.

FIGURE 3.16: Table of Contents (2) of Test5 Document



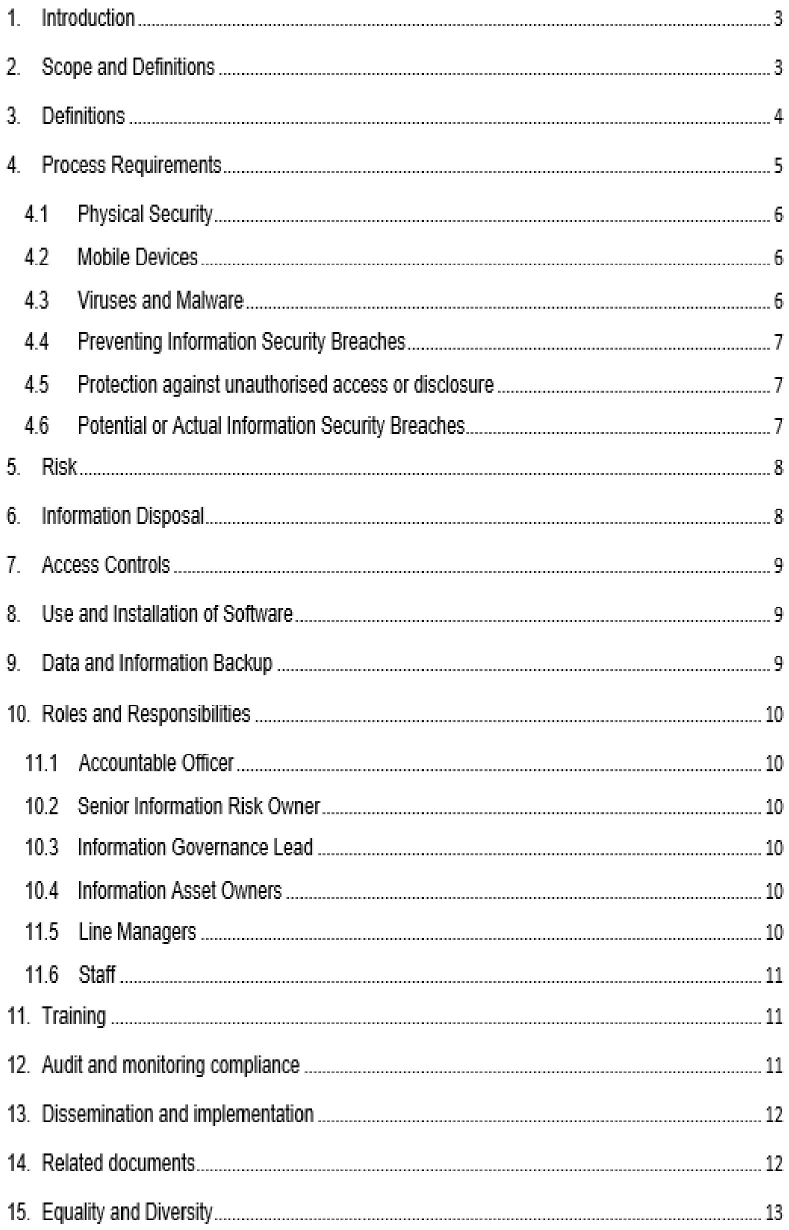
Personal Information (PI)	It has the meaning set out in section 2 of the Freedom of Information and Protection of Privacy Act (FIPPA) as: recorded Information about an identifiable individual, including, a. Information relating to the race, national or ethnic origin, color, religion, age, sex, sexual orientation or marital or family status of the individual, b. Information relating to the education or the medical, psychiatric, psychological, criminal or employment history of the individual or Information relating to financial transactions in which the individual has been involved, c. any identifying number, symbol or other particular assigned to the individual, d. the address, telephone number, fingerprints or blood type of the individual, e. the personal opinions or views of the individual except where they relate to another individual, f. correspondence sent to an institution by the individual that is implicitly or explicitly of a private or confidential nature, and replies to that correspondence that would reveal the contents of the original correspondence, g. the views or opinions of another individual about the individual, and h. the individual's name where it appears with other Personal Information relating to the individual or where the disclosure of the name would reveal other Personal Information about the individual.
Risk Treatment	Management's decision to manage the risk, (transfer, avoid, mitigate, accept) and action(s) that may be taken to bring the risk situation to a level where the exposure to risk is acceptable to eHealth Ontario based on risk appetite.
Safeguard	A precautionary measure, stipulation, device, technical or non-technical solution to prevent an undesired incident from occurring.
Security Incident	Any activity that could compromise the security of Information or systems, including but not limited to, a social engineering attempt such as a request for a password, loss of a laptop or blackberry, a computer virus infection, degradation of a system, unauthorized changes to files or file sizes, or the addition of files.
Security Posture	The security status of an enterprise's networks, information, and systems based on information security resources (e.g., people, hardware, software, policies) and capabilities in place to manage the defense of the enterprise and to react as the situation changes.
Segregation of Duties	Principle of having more than one person required to complete a specific task. This process is a control used to prevent fraud and error.
Sensitive Information	Information that if released without authorization would cause harm, embarrassment, or unfair economic advantage, i.e., a breach of confidentiality of Personal Information, Personal Health Information, unauthorized modification of financial data, or a release of pre-budget information and strategic planning documents.

Figure 3.17: Table of Contents (3) of Test5 Document

Test 6 [17]**:** IT security policy document of Mid Essex Hospital Services NHS Trust

- Last reviewed in 2018
- Total number of pages: 16

[17] http://stellarhealthcare.net/images/policies/Information_Security_Policy_Stellar_Healthcare_v1.0_Final.doc



1. Introduction 3

2. Scope and Definitions 3

3. Definitions 4

4. Process Requirements 5

4.1 Physical Security 6

4.2 Mobile Devices 6

4.3 Viruses and Malware 6

4.4 Preventing Information Security Breaches 7

4.5 Protection against unauthorised access or disclosure 7

4.6 Potential or Actual Information Security Breaches 7

5. Risk 8

6. Information Disposal 8

7. Access Controls 9

8. Use and Installation of Software 9

9. Data and Information Backup 9

10. Roles and Responsibilities 10

11.1 Accountable Officer 10

10.2 Senior Information Risk Owner 10

10.3 Information Governance Lead 10

10.4 Information Asset Owners 10

11.5 Line Managers 10

11.6 Staff 11

11. Training 11

12. Audit and monitoring compliance 11

13. Dissemination and implementation 12

14. Related documents 12

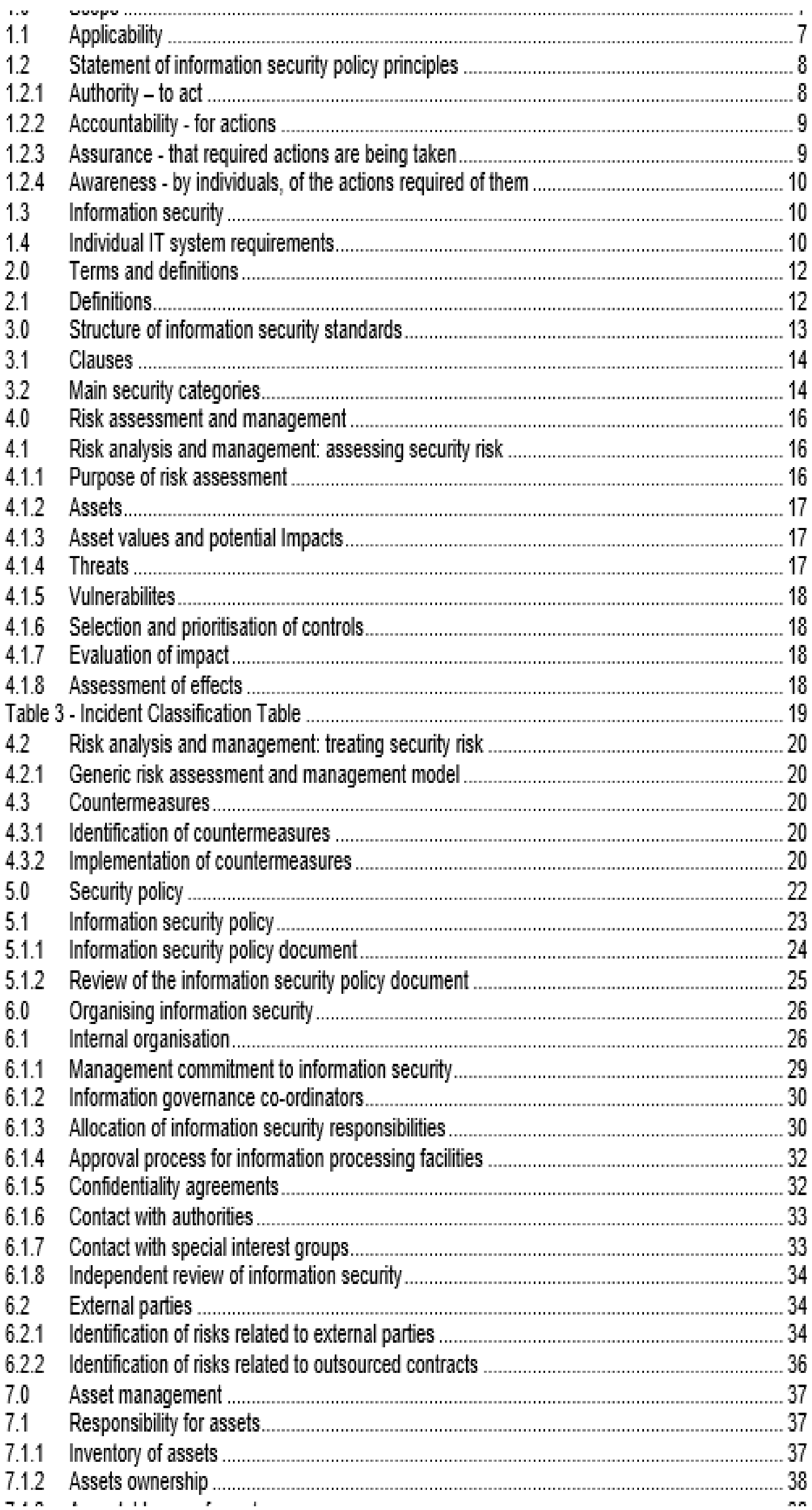
15. Equality and Diversity 13

Figure 3.18: Table of Contents of Test6 Document

Test 7 [18]**:** IT security policy document of NHS Scotland.

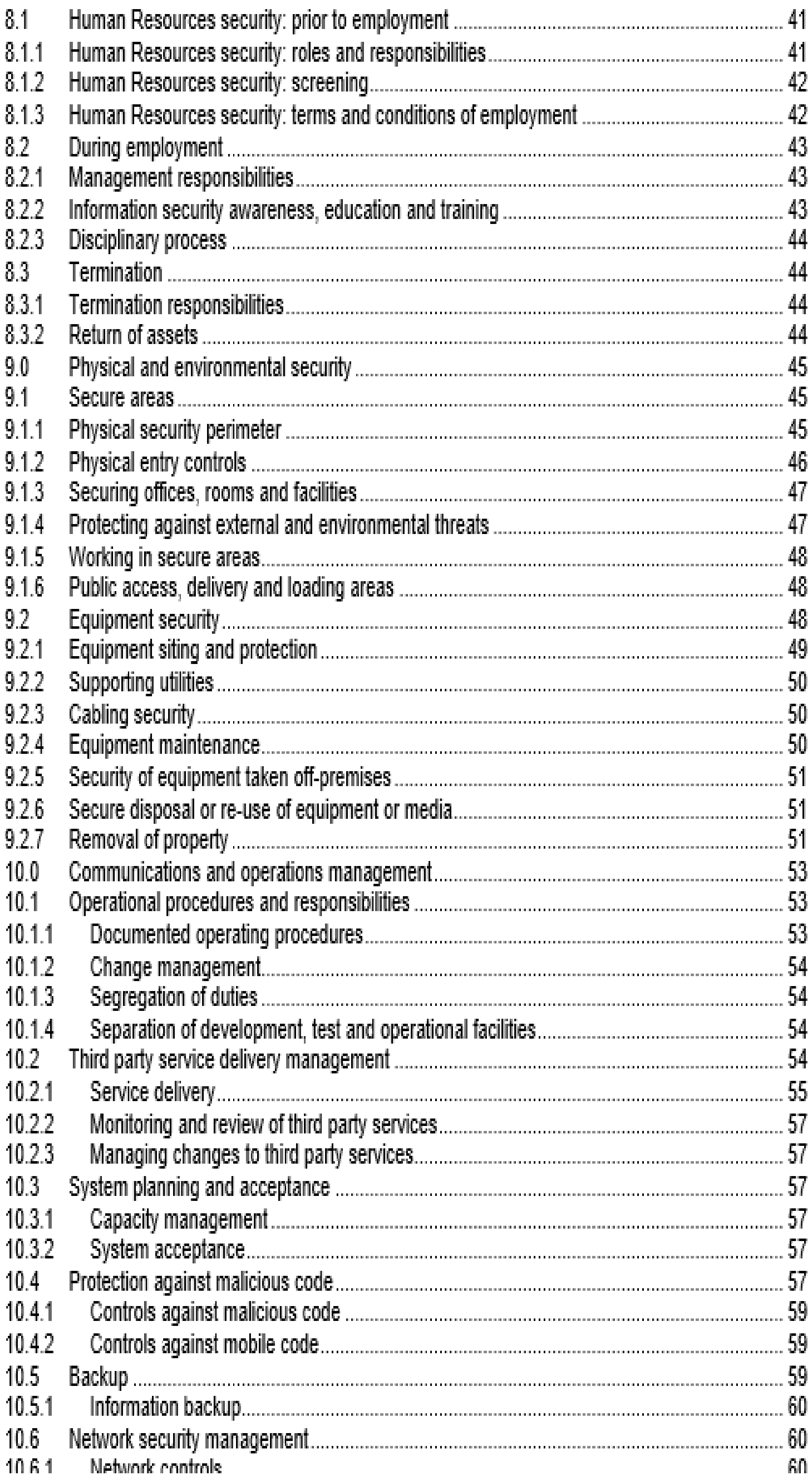
- Based on ISO17799
- Last reviewed in 2005
- Total number of pages: 108

[18] https://www.ehealth.scot/wp-content/uploads/documents/standard-security-policy-and-standards.doc#: :text=It%20is%20the%20Policy%20of,legislative%20requirements%20will%20be%20 assured.&text=Information%20security%20training%20will%20be%20available%20to%20all%20staff.



1.0 Scope 7
1.1 Applicability 7
1.2 Statement of information security policy principles 8
1.2.1 Authority – to act 8
1.2.2 Accountability - for actions 9
1.2.3 Assurance - that required actions are being taken 9
1.2.4 Awareness - by individuals, of the actions required of them 10
1.3 Information security 10
1.4 Individual IT system requirements 10
2.0 Terms and definitions 12
2.1 Definitions 12
3.0 Structure of information security standards 13
3.1 Clauses 14
3.2 Main security categories 14
4.0 Risk assessment and management 16
4.1 Risk analysis and management: assessing security risk 16
4.1.1 Purpose of risk assessment 16
4.1.2 Assets 17
4.1.3 Asset values and potential Impacts 17
4.1.4 Threats 17
4.1.5 Vulnerabilites 18
4.1.6 Selection and prioritisation of controls 18
4.1.7 Evaluation of impact 18
4.1.8 Assessment of effects 18
Table 3 - Incident Classification Table 19
4.2 Risk analysis and management: treating security risk 20
4.2.1 Generic risk assessment and management model 20
4.3 Countermeasures 20
4.3.1 Identification of countermeasures 20
4.3.2 Implementation of countermeasures 20
5.0 Security policy 22
5.1 Information security policy 23
5.1.1 Information security policy document 24
5.1.2 Review of the information security policy document 25
6.0 Organising information security 26
6.1 Internal organisation 26
6.1.1 Management commitment to information security 29
6.1.2 Information governance co-ordinators 30
6.1.3 Allocation of information security responsibilities 30
6.1.4 Approval process for information processing facilities 32
6.1.5 Confidentiality agreements 32
6.1.6 Contact with authorities 33
6.1.7 Contact with special interest groups 33
6.1.8 Independent review of information security 34
6.2 External parties 34
6.2.1 Identification of risks related to external parties 34
6.2.2 Identification of risks related to outsourced contracts 36
7.0 Asset management 37
7.1 Responsibility for assets 37
7.1.1 Inventory of assets 37
7.1.2 Assets ownership 38

Figure 3.19: Table of Contents (1) of Test 7 Document



8.1 Human Resources security: prior to employment 41
8.1.1 Human Resources security: roles and responsibilities 41
8.1.2 Human Resources security: screening 42
8.1.3 Human Resources security: terms and conditions of employment 42
8.2 During employment 43
8.2.1 Management responsibilities 43
8.2.2 Information security awareness, education and training 43
8.2.3 Disciplinary process 44
8.3 Termination 44
8.3.1 Termination responsibilities 44
8.3.2 Return of assets 44
9.0 Physical and environmental security 45
9.1 Secure areas 45
9.1.1 Physical security perimeter 45
9.1.2 Physical entry controls 46
9.1.3 Securing offices, rooms and facilities 47
9.1.4 Protecting against external and environmental threats 47
9.1.5 Working in secure areas 48
9.1.6 Public access, delivery and loading areas 48
9.2 Equipment security 48
9.2.1 Equipment siting and protection 49
9.2.2 Supporting utilities 50
9.2.3 Cabling security 50
9.2.4 Equipment maintenance 50
9.2.5 Security of equipment taken off-premises 51
9.2.6 Secure disposal or re-use of equipment or media 51
9.2.7 Removal of property 51
10.0 Communications and operations management 53
10.1 Operational procedures and responsibilities 53
10.1.1 Documented operating procedures 53
10.1.2 Change management 54
10.1.3 Segregation of duties 54
10.1.4 Separation of development, test and operational facilities 54
10.2 Third party service delivery management 54
10.2.1 Service delivery 55
10.2.2 Monitoring and review of third party services 57
10.2.3 Managing changes to third party services 57
10.3 System planning and acceptance 57
10.3.1 Capacity management 57
10.3.2 System acceptance 57
10.4 Protection against malicious code 57
10.4.1 Controls against malicious code 59
10.4.2 Controls against mobile code 59
10.5 Backup 59
10.5.1 Information backup 60
10.6 Network security management 60
10.6.1 Network controls 60

Figure 3.20: Table of Contents (2) of Test 7 Document

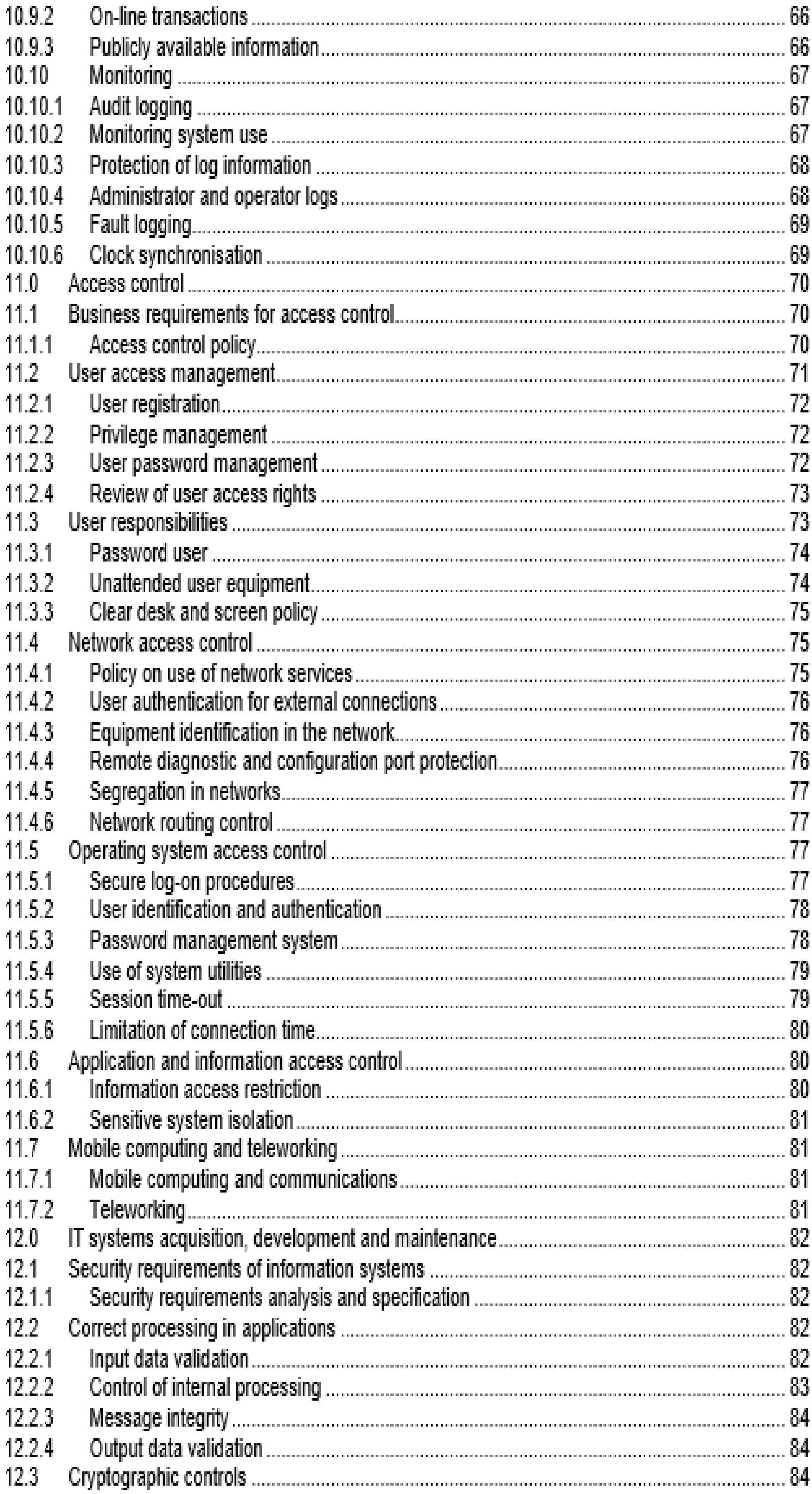
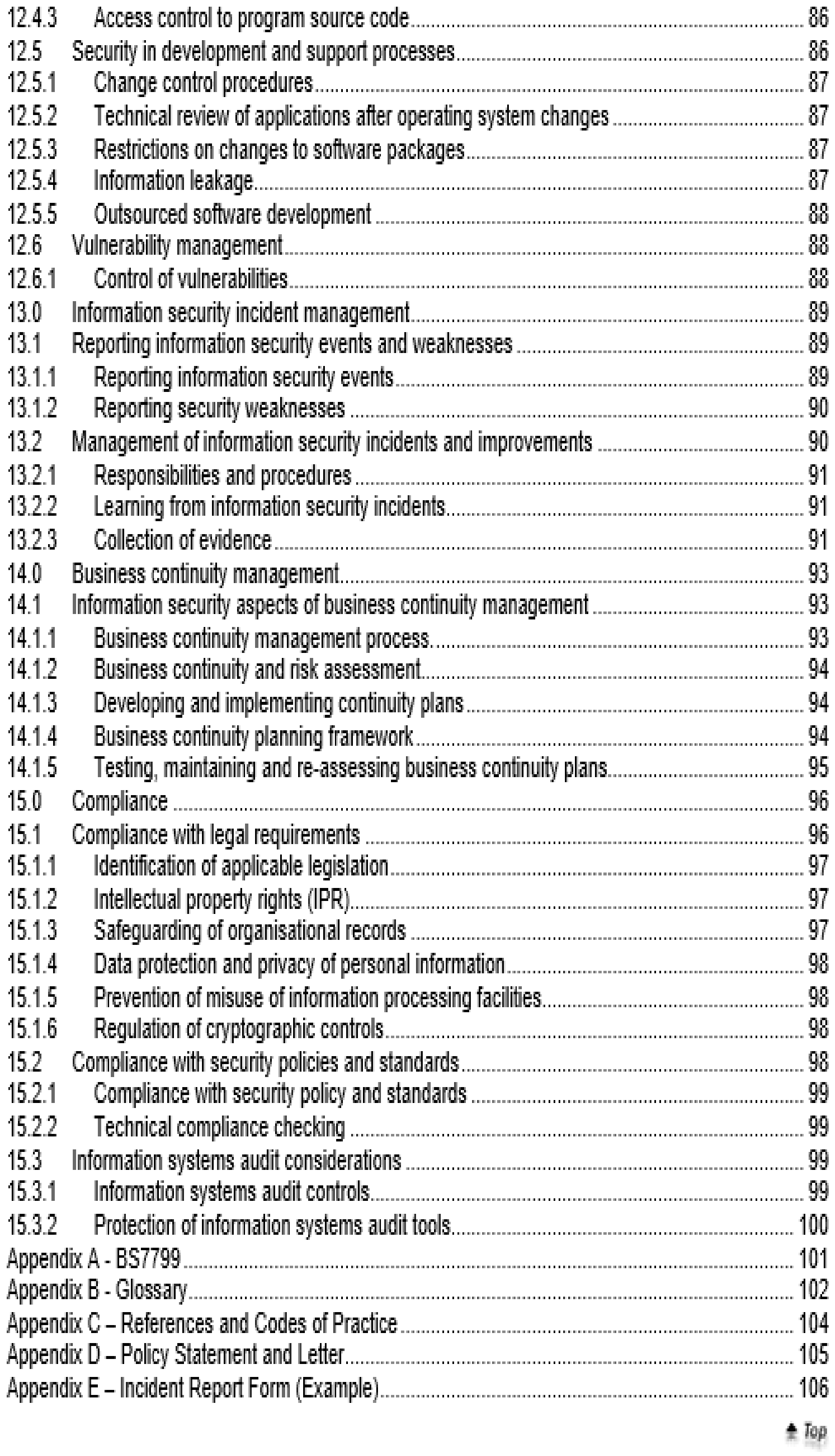


Figure 3.21: Table of Contents (3) of Test 7 Document

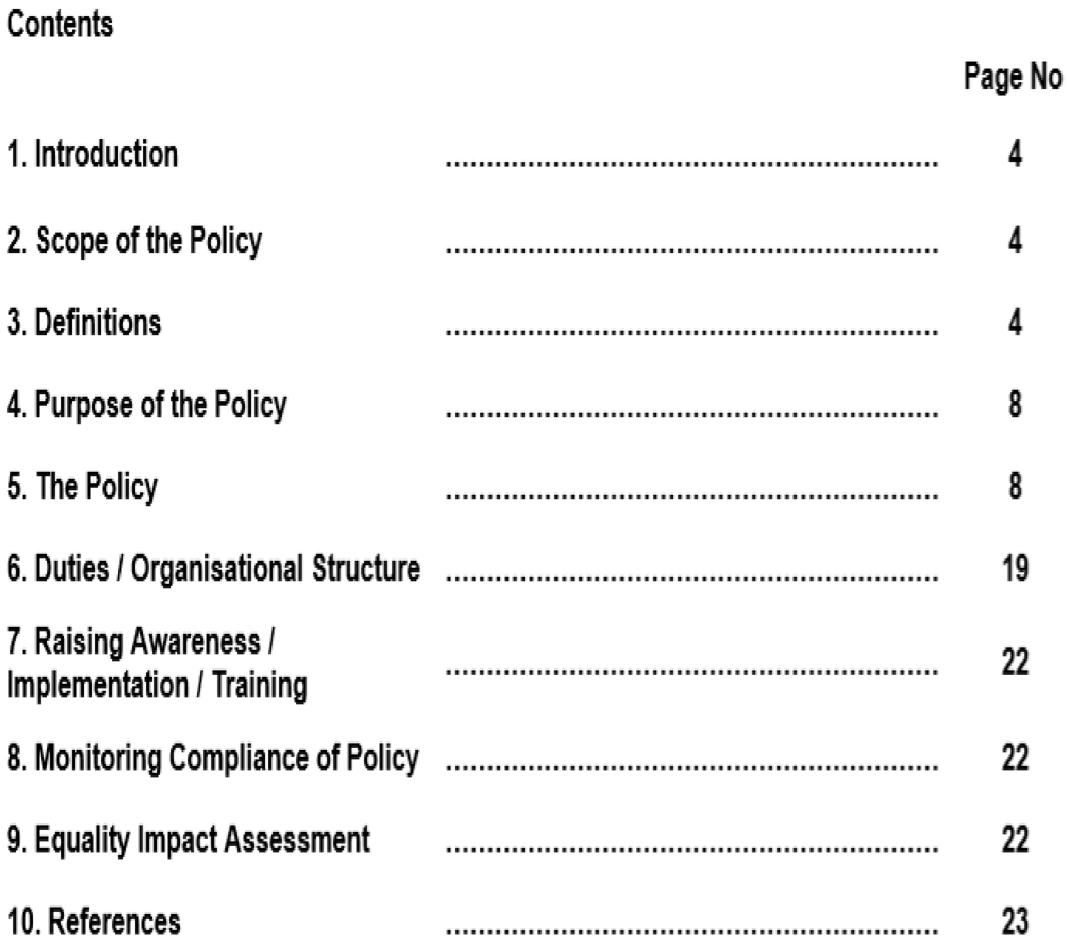


12.4.3 Access control to program source code 86
12.5 Security in development and support processes 86
12.5.1 Change control procedures 87
12.5.2 Technical review of applications after operating system changes 87
12.5.3 Restrictions on changes to software packages 87
12.5.4 Information leakage 87
12.5.5 Outsourced software development 88
12.6 Vulnerability management 88
12.6.1 Control of vulnerabilities 88
13.0 Information security incident management 89
13.1 Reporting information security events and weaknesses 89
13.1.1 Reporting information security events 89
13.1.2 Reporting security weaknesses 90
13.2 Management of information security incidents and improvements 90
13.2.1 Responsibilities and procedures 91
13.2.2 Learning from information security incidents 91
13.2.3 Collection of evidence 91
14.0 Business continuity management 93
14.1 Information security aspects of business continuity management 93
14.1.1 Business continuity management process 93
14.1.2 Business continuity and risk assessment 94
14.1.3 Developing and implementing continuity plans 94
14.1.4 Business continuity planning framework 94
14.1.5 Testing, maintaining and re-assessing business continuity plans 95
15.0 Compliance 96
15.1 Compliance with legal requirements 96
15.1.1 Identification of applicable legislation 97
15.1.2 Intellectual property rights (IPR) 97
15.1.3 Safeguarding of organisational records 97
15.1.4 Data protection and privacy of personal information 98
15.1.5 Prevention of misuse of information processing facilities 98
15.1.6 Regulation of cryptographic controls 98
15.2 Compliance with security policies and standards 98
15.2.1 Compliance with security policy and standards 99
15.2.2 Technical compliance checking 99
15.3 Information systems audit considerations 99
15.3.1 Information systems audit controls 99
15.3.2 Protection of information systems audit tools 100
Appendix A - BS7799 101
Appendix B - Glossary 102
Appendix C – References and Codes of Practice 104
Appendix D – Policy Statement and Letter 105
Appendix E – Incident Report Form (Example) 106

FIGURE 3.22: Table of Contents (4) of Test 7 Document

Test 8 [19]**:** IT security policy document of Frimley Health Foundation.

- Last reviewed in 2020
- Total number of pages: 23



Contents

	Page No
1. Introduction	4
2. Scope of the Policy	4
3. Definitions	4
4. Purpose of the Policy	8
5. The Policy	8
6. Duties / Organisational Structure	19
7. Raising Awareness / Implementation / Training	22
8. Monitoring Compliance of Policy	22
9. Equality Impact Assessment	22
10. References	23

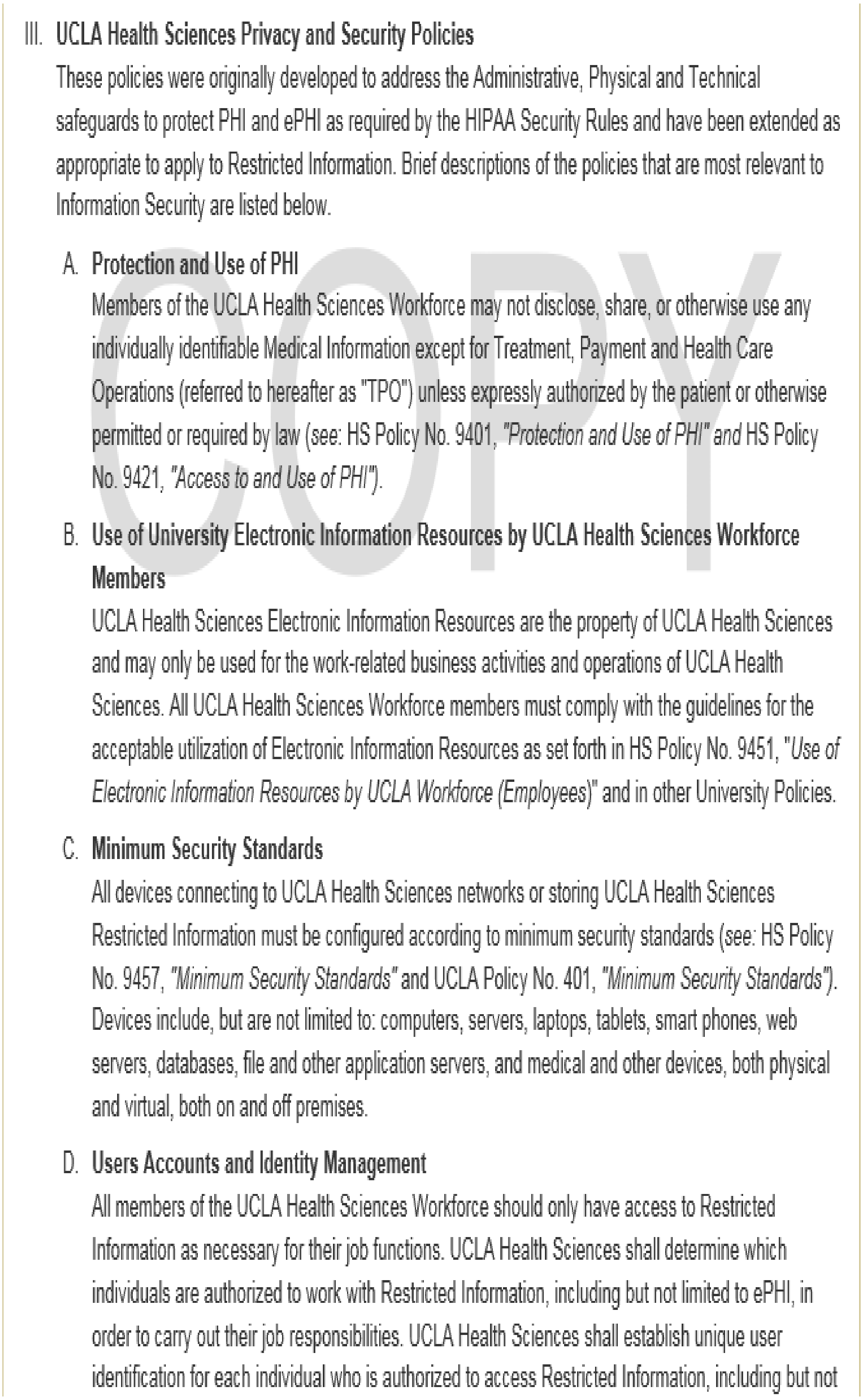
Figure 3.23: Table of Contents of Test 8 Document

Test 9 [20]**:** IT security policy document of UCLA Medical Center.

- Last reviewed in 2020
- Total number of pages: 7

[19] https://www.uclahealth.org/compliance/workfiles/HS%20Policies/HS9450-InformationSecurity.pdf

[20] https://www.fhft.nhs.uk/media/4234/information-security-policy.pdf



III. **UCLA Health Sciences Privacy and Security Policies**

These policies were originally developed to address the Administrative, Physical and Technical safeguards to protect PHI and ePHI as required by the HIPAA Security Rules and have been extended as appropriate to apply to Restricted Information. Brief descriptions of the policies that are most relevant to Information Security are listed below.

A. **Protection and Use of PHI**

Members of the UCLA Health Sciences Workforce may not disclose, share, or otherwise use any individually identifiable Medical Information except for Treatment, Payment and Health Care Operations (referred to hereafter as "TPO") unless expressly authorized by the patient or otherwise permitted or required by law (*see:* HS Policy No. 9401, *"Protection and Use of PHI" and* HS Policy No. 9421, *"Access to and Use of PHI"*).

B. **Use of University Electronic Information Resources by UCLA Health Sciences Workforce Members**

UCLA Health Sciences Electronic Information Resources are the property of UCLA Health Sciences and may only be used for the work-related business activities and operations of UCLA Health Sciences. All UCLA Health Sciences Workforce members must comply with the guidelines for the acceptable utilization of Electronic Information Resources as set forth in HS Policy No. 9451, "*Use of Electronic Information Resources by UCLA Workforce (Employees)*" and in other University Policies.

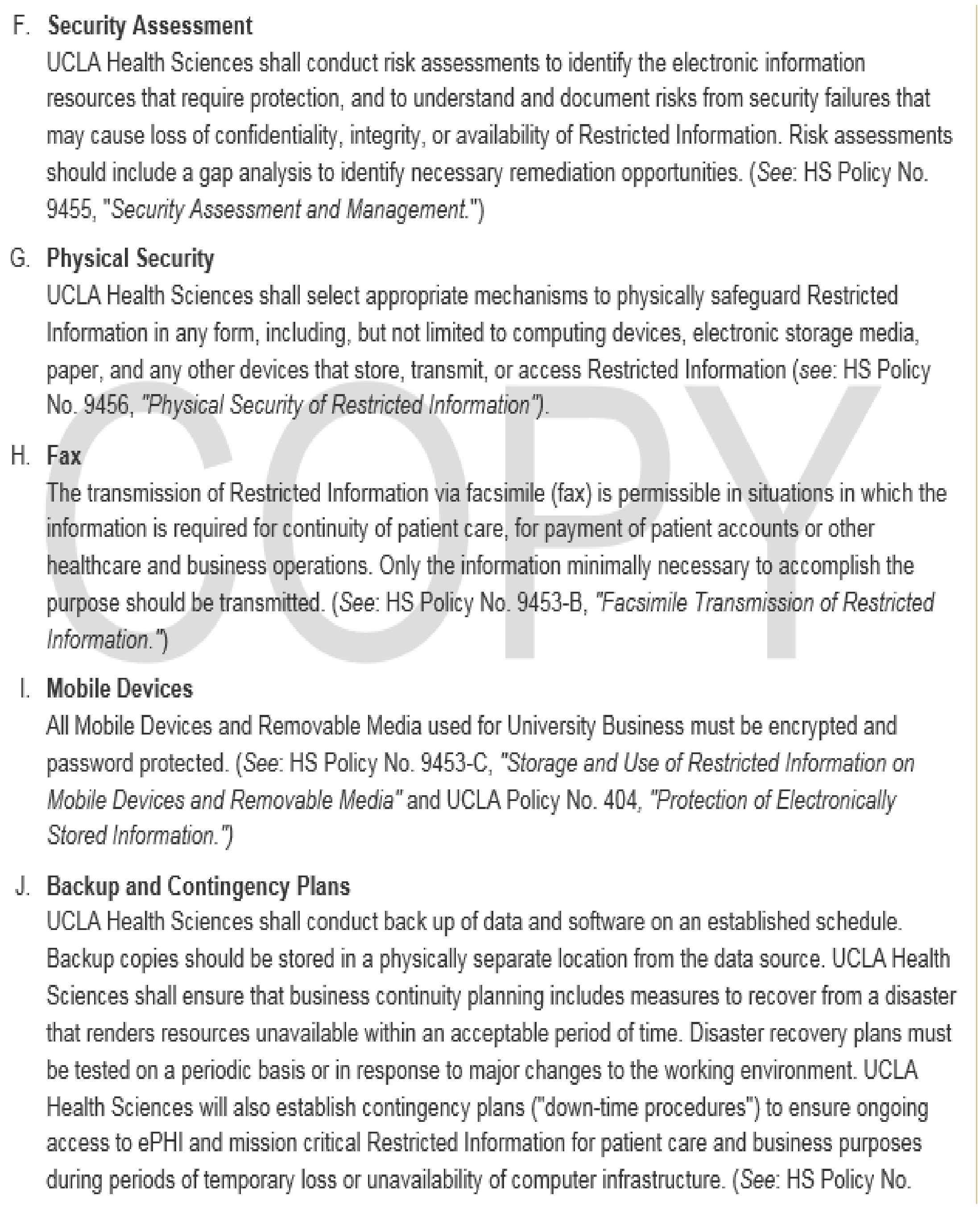
C. **Minimum Security Standards**

All devices connecting to UCLA Health Sciences networks or storing UCLA Health Sciences Restricted Information must be configured according to minimum security standards (*see:* HS Policy No. 9457, *"Minimum Security Standards"* and UCLA Policy No. 401, *"Minimum Security Standards"*). Devices include, but are not limited to: computers, servers, laptops, tablets, smart phones, web servers, databases, file and other application servers, and medical and other devices, both physical and virtual, both on and off premises.

D. **Users Accounts and Identity Management**

All members of the UCLA Health Sciences Workforce should only have access to Restricted Information as necessary for their job functions. UCLA Health Sciences shall determine which individuals are authorized to work with Restricted Information, including but not limited to ePHI, in order to carry out their job responsibilities. UCLA Health Sciences shall establish unique user identification for each individual who is authorized to access Restricted Information, including but not

FIGURE 3.24: Table of Contents (1) of Test 9 Document



F. **Security Assessment**
UCLA Health Sciences shall conduct risk assessments to identify the electronic information resources that require protection, and to understand and document risks from security failures that may cause loss of confidentiality, integrity, or availability of Restricted Information. Risk assessments should include a gap analysis to identify necessary remediation opportunities. (*See*: HS Policy No. 9455, "*Security Assessment and Management.*")

G. **Physical Security**
UCLA Health Sciences shall select appropriate mechanisms to physically safeguard Restricted Information in any form, including, but not limited to computing devices, electronic storage media, paper, and any other devices that store, transmit, or access Restricted Information (*see*: HS Policy No. 9456, *"Physical Security of Restricted Information"*).

H. **Fax**
The transmission of Restricted Information via facsimile (fax) is permissible in situations in which the information is required for continuity of patient care, for payment of patient accounts or other healthcare and business operations. Only the information minimally necessary to accomplish the purpose should be transmitted. (*See*: HS Policy No. 9453-B, *"Facsimile Transmission of Restricted Information."*)

I. **Mobile Devices**
All Mobile Devices and Removable Media used for University Business must be encrypted and password protected. (*See*: HS Policy No. 9453-C, *"Storage and Use of Restricted Information on Mobile Devices and Removable Media"* and UCLA Policy No. 404, *"Protection of Electronically Stored Information."*)

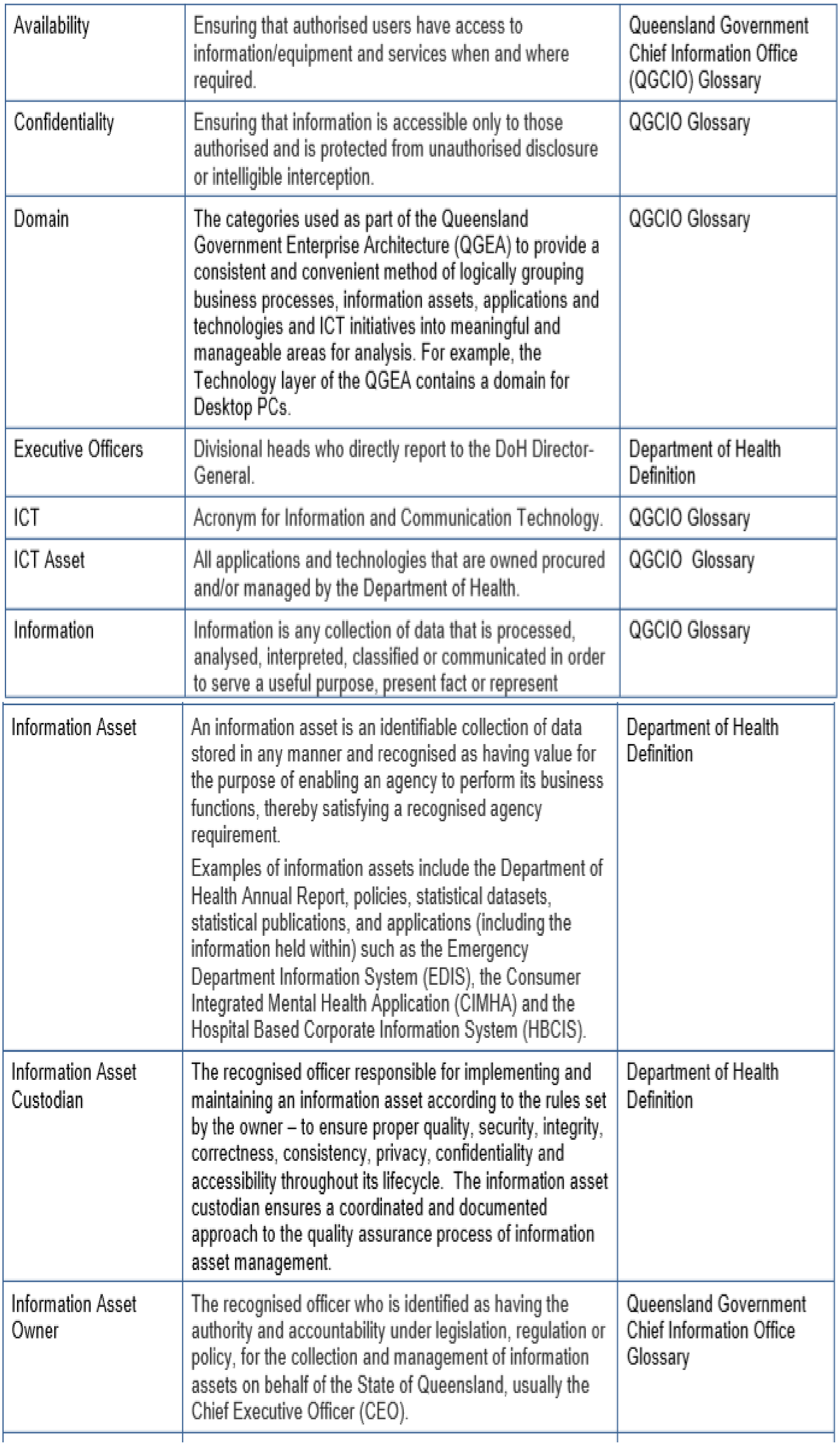
J. **Backup and Contingency Plans**
UCLA Health Sciences shall conduct back up of data and software on an established schedule. Backup copies should be stored in a physically separate location from the data source. UCLA Health Sciences shall ensure that business continuity planning includes measures to recover from a disaster that renders resources unavailable within an acceptable period of time. Disaster recovery plans must be tested on a periodic basis or in response to major changes to the working environment. UCLA Health Sciences will also establish contingency plans ("down-time procedures") to ensure ongoing access to ePHI and mission critical Restricted Information for patient care and business purposes during periods of temporary loss or unavailability of computer infrastructure. (*See*: HS Policy No.

Figure 3.25: Table of Contents (2) of Test 9 Document

Test 10 [21]**:** IT security policy document of Queensland Hospital.

- Last reviewed in 2014
- Total number of pages: 3

[21] https://www.health.qld.gov.au/_data/assets/pdf_file/0041/859595/qh-pol-468.pdf



Availability	Ensuring that authorised users have access to information/equipment and services when and where required.	Queensland Government Chief Information Office (QGCIO) Glossary
Confidentiality	Ensuring that information is accessible only to those authorised and is protected from unauthorised disclosure or intelligible interception.	QGCIO Glossary
Domain	The categories used as part of the Queensland Government Enterprise Architecture (QGEA) to provide a consistent and convenient method of logically grouping business processes, information assets, applications and technologies and ICT initiatives into meaningful and manageable areas for analysis. For example, the Technology layer of the QGEA contains a domain for Desktop PCs.	QGCIO Glossary
Executive Officers	Divisional heads who directly report to the DoH Director-General.	Department of Health Definition
ICT	Acronym for Information and Communication Technology.	QGCIO Glossary
ICT Asset	All applications and technologies that are owned procured and/or managed by the Department of Health.	QGCIO Glossary
Information	Information is any collection of data that is processed, analysed, interpreted, classified or communicated in order to serve a useful purpose, present fact or represent	QGCIO Glossary
Information Asset	An information asset is an identifiable collection of data stored in any manner and recognised as having value for the purpose of enabling an agency to perform its business functions, thereby satisfying a recognised agency requirement. Examples of information assets include the Department of Health Annual Report, policies, statistical datasets, statistical publications, and applications (including the information held within) such as the Emergency Department Information System (EDIS), the Consumer Integrated Mental Health Application (CIMHA) and the Hospital Based Corporate Information System (HBCIS).	Department of Health Definition
Information Asset Custodian	The recognised officer responsible for implementing and maintaining an information asset according to the rules set by the owner – to ensure proper quality, security, integrity, correctness, consistency, privacy, confidentiality and accessibility throughout its lifecycle. The information asset custodian ensures a coordinated and documented approach to the quality assurance process of information asset management.	Department of Health Definition
Information Asset Owner	The recognised officer who is identified as having the authority and accountability under legislation, regulation or policy, for the collection and management of information assets on behalf of the State of Queensland, usually the Chief Executive Officer (CEO).	Queensland Government Chief Information Office Glossary

Figure 3.26: Table of Contents of Test 10 Document

3.4 Proposed System Architecture

The similarity score is calculated with the help of three different techniques; Cosine Similarity Measure, Jaccard Similarity Measure and String-based Similarity Measure. The output of this experiment is the similarity score of test document as compared to the standard document. This similarity score is the weighted sum of all three above mentioned techniques.

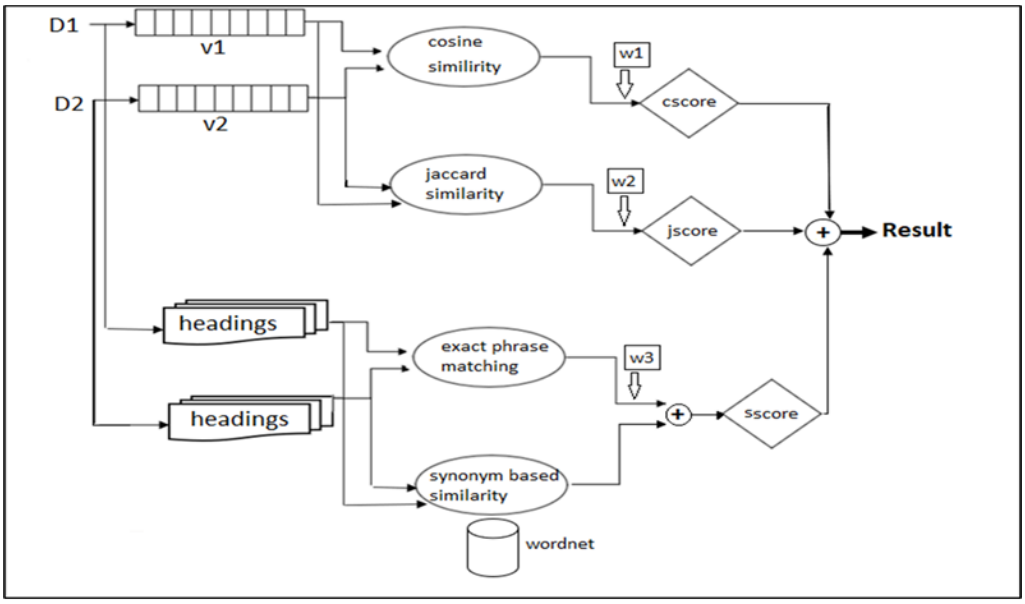


FIGURE 3.27: System Architecture Diagram

The architecture diagram shows the flow of the proposed technique. In the first step, text is extracted from the documents. This solution is scalable. It can be given any number of documents and it will compute the expected results in the same way. After the preprocessing, vectors are created from the extracted text.

These vectors are taken as input to calculate similarity scores from cosine measure and Jaccard measure separately. The headings from the documents are extracted with the help of python library python-docx.

Firstly, these headings are compared phrase by phrase. Then, with the help of WordNet dictionary, headings are compared on the basis of synonyms or synsets. An average of scores of exact phrase matching and synonym-based similarity is calculated.

Before combining the scores from cosine similarity, jaccard similarity and string similarity, these scores are assigned weights in order to normalize the result. String similarity measure is assigned the maximum weight, i.e., 0.6, while both of the other two similarity measures are assigned a weight of 0.2 each. The weighted sum of all three techniques is the final similarity score.

3.4.1 Structural Similarity Calculation

The rhetoric or structural similarity between the documents is calculated by ontology-based comparison of the headings of the documents, by exact phrase matching and also lexically, so that the similarity score is not affected even if the headings have different words but they are similar in meaning. Headings from the documents are extracted using the python library, python-docx. For the word-to-word analysis, string matching is used.

WordNet library is a lexical database of semantic relations between words in natural languages. It is used to check if the content of headings is similar in meaning even if they dont match as strings. The advantage of using WordNet is that it contains words and relationships that are highly accurate, because it was manually constructed.

3.4.2 Content-Based Similarity Calculation

Content-based similarity between the documents is calculated with the help of vector space-based count similarity metrices, i.e., Cosine Similarity Measure and Jaccard Similarity Measure. Cosine Similarity is calculated by finding out the angle between two vectors. When both vectors are equal, the cosine similarity index is 1, when both vectors are perpendicular, the cosine similarity index is 0 and it is -1 when vectors are completely opposite. Cosine similarity is basically the angle of deviation of one vector from the other. The reason for using this measure is that it calculates the similarity based on the direction of the vector rather than its magnitude, which means that the comparison of two documents will give efficient results even if they are very different in lengths. Jaccard similarity is calculating by dividing size of intersection over size of union. Its value varies between 0 and 1. Both of these techniques are used simultaneously instead of choosing one of them to get accurate results as Jaccard similarity is more suitable for cases where duplication of words does not matter but cosine similarity takes duplication in account calculating text similarity. A weighted sum of all of the above-mentioned techniques gives the final similarity score.

3.5 Proposed Solution

The proposed solution is calculated on the basis of scores from all three techniques mentioned in the system architecture. The main steps involved for the calculation of similarity score are as follows:

1. Extract text from documents
2. Preprocessing of the text
3. Text Vectorization
4. Calculation of string-based similarity
5. Calculation of Cosine Similarity
6. Calculation of Jaccard Similarity
7. Calculation of weighted sum-based similarity calculation

3.5.1 Extract Text from Documents

As the documents under consideration are in word format, extracting text from them was very easy using the Python libraries like python-docx and PyPDF2. The pre-installed Python libraries helped to separate headings from the text.

3.5.2 Preprocessing of the Text

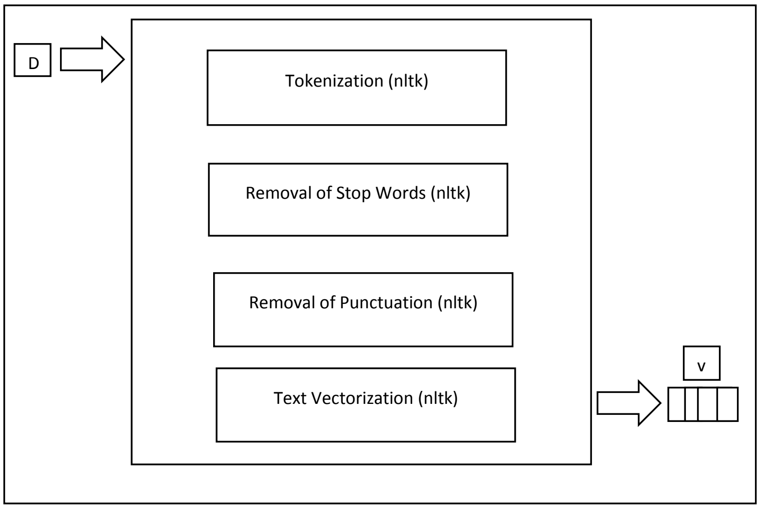

D
Tokenization (nltk)
Removal of Stop Words (nltk)
Removal of Punctuation (nltk)
Text Vectorization (nltk)
V

Figure 3.28: Preprocessing

Once the text is extracted and discerned, the second step is the preprocessing. Preprocessing can be defined as filtering the text and bringing it in such a form which is easier to manipulate. It consists of various steps like removal of unwanted words, removal of punctuation and tokenization. Tokenization refers to dividing the text into smaller one- or two-word parts. The tokenizer punkt of nltk is used for this purpose.

3.5.3 Text Vectorization

The purpose of text vectorization in natural language processing is to characterize the text into numerical form so that its manipulation becomes easier. In this process, each token is mapped to a corresponding vector of real number. There are various approaches of text vectorization which are used while analyzing text similarity.

3.5.4 Calculation of String-Based Similarity

The string-based similarity score is calculated in two steps. First the headings of the test document are compared to the headings of standard documents by exact phrase matching. In the next step, both phrases are again compared to each other on the basis of their synonyms. This is done with the help of the wordnet library of the nltk package. If a string does not match the other string but its synonym is found in the second string, it can be detected in this step. In the final step, an average score is calculated by adding both of the previously calculated scores. Along with word-to-word comparisons, the meanings of both strings are also checked for similarity to get more accurate results.

3.5.5 Calculation of Cosine Similarity

Cosine similarity measure is calculated irrespective of the size of documents because it focuses on the angle between two vectors, so even if both documents have different sizes, this measure will still give accurate results. The score calculated by this measure will have a value between 0 to 1. The mathematical formula found in literature for the calculation of cosine similarity is as follows:

$$\text{Cosine Similarity}\,(\text{A},\ \text{B}) = \frac{A.B}{||\text{A}|\ \times\ |\text{B}||} \tag{3.1}$$

$$\cos\theta\ = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \times \sqrt{\sum_1^n b_i^2}} \tag{3.2}$$

Where a.b is the dot product of the two vectors.

3.5.6 Calculation of Jaccard Similarity

Jaccard Similarity is defined as the size of the intersection divided by the size of the union of two sets. It is used to calculate the similarity between any two objects, in this case, documents. Similarity is measured by dividing the intersection of the items by the union of items.

The score calculated by this measure will have a value between 0 to 1. The value 1 means that both documents have maximum similarity A=B and the value 0 means that both documents are disjoint and completely different. The mathematical formula found in literature for the calculation of Jaccard similarity is as follows:

$$Jaccard\ Similarity\ (A,B)\ =\ \frac{|A\cap B|}{|A|+|B|-|A\cap B|}\ =\ \frac{|A\cap B|}{|A\cup B|} \tag{3.3}$$

3.5.7 Weighted Sum-Based Similarity Calculation

The scores calculated from cosine similarity measure, Jaccard similarity measure and string similarity measure are then assigned some weights. The total sum of these weights is 1. String similarity measure is assigned the maximum weight, i.e., 0.6, while both of the other two similarity measures are assigned a weight of 0.2 each. Proposed formula for calculation of Similarity Score of each document:

$$\mathrm{Sim(D1,D2) = Csim + Jsim + Ssim} \tag{3.4}$$

$$Csim =\ CosineSim(D1,D2)\times W1 \tag{3.5}$$

$$Ssim =\ StringSim(D1,D2)\times W3 \tag{3.6}$$

$$\text{Where; w3 < w1 \& w2} \tag{3.7}$$

In this experiment, string-based similarity is being calculated solely on the basis of headings of the document, i.e., structure of the document. As the size of input is increased, the noise in results also increases because of irrelevant terms. Overall content may contain some irrelevant terms which can hinder the results but the chance of noise is very low in case of exact phrase matching of the headings. For this reason, string-based similarity measure is given the maximum weight.

The weighted sum computed at the end of this experiment is the final result. A similarity score is calculated against every test document and the results are compared in the next chapter.The results are evaluated on the basis of user-based study. The four measures used for user-based evaluation are accuracy, precision, recall and F-Measure.

Chapter 4

Results and Discussion

This chapter provides thorough discussion on the experiment and results achieved by implementation of the methodology discussed in the previous chapter. Moreover, comparison of similarity techniques with proposed technique is also presented in the chapter.

4.1 Data Collection

The data used in this experiment is in the form of word documents. In order to choose the right documents for the study, exhaustive web search is performed. Information Security Policy Documents from various domains are available on the web but only a single domain is focused in order to get more precise results. The chosen domain for this experiment is health sector. Security in health organizations is a topic of growing interest. Many health organizations have just started learning and adopting security procedures. The security policy documents from authentic sources of healthcare are easily available. Out of many such documents, only 11 documents are shortlisted after thoroughly reviewing them.

The documents that are finally chosen for the experiment are collected from the websites of National Learning Consortium (NLC) [1], NHS Scotland [2], Portsmouth

[1] https://www.healthit.gov/

[2] https://www.scot.nhs.uk/

Hospital [3], Stellar Healthcare Organization [4] and Swiss Personalized Health Network [5], etc.

4.2 Documents Preprocessing and Data Extraction

Most of the documents found were in PDF or word format. Python libraries can be used to easily extract different types of data from a single file. Out of 11 documents, 8 documents were already in word format while 3 documents were in pdf format. In order for data to be uniform, only word (.docx) documents were used.

Three documents that were not already in word format are be converted from PDF to words with the help of the tool PDFtodocx. The Python libraries used for the extraction of text from the documents are python-docx and PyPDF2.

4.3 Text Preprocessing

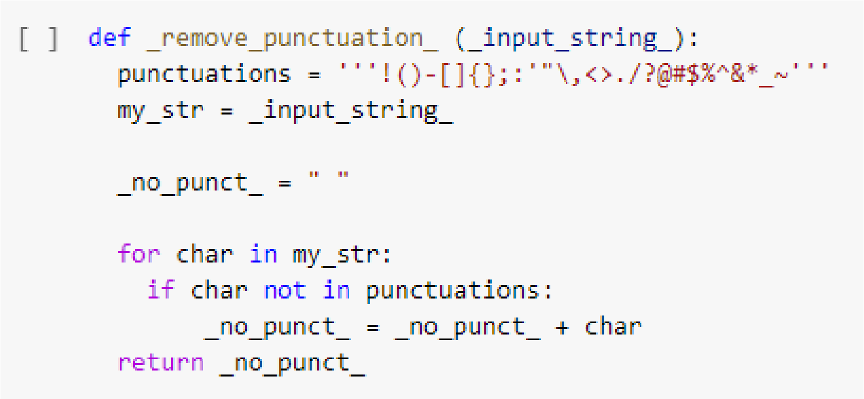
The text extracted from the documents is further preprocessed to filter out any noise in it in order to get more accurate results. The preprocessing of text consists of 4 steps. At first, the text is tokenized using built-in tokenizer of the nltk library.

The tokenizer divides the text into smaller chunks which are easier to manipulate. After the tokenization, useless words, which might hinder the overall result, from the text are removed. These words are referred to as stop words e.g., ”is”, ”an”, ”the”, etc. In the 3rd step, punctuation marks from the text are removed. The filtered tokens are then converted into vectors for further processing. Every token is mapped to a number during text data vectorization. Example code is given below.

[3] https://www.porthosp.nhs.uk/

[4] http://www.stellarhealthcare.net/

[5] https://sphn.ch/



```
[ ]  def _remove_punctuation_ (_input_string_):
       punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
       my_str = _input_string_

       _no_punct_ = " "

       for char in my_str:
         if char not in punctuations:
             _no_punct_ = _no_punct_ + char
       return _no_punct_
```

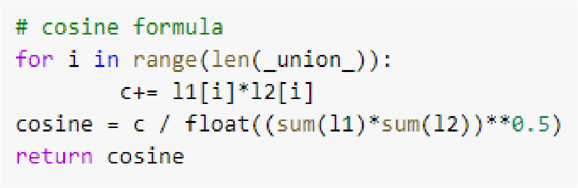
Figure 4.1: Preprocessing; removal of punctuation

4.4 Similarity Techniques

3 different similarity techniques are used in this experiment. The new proposed technique is the weighted sum of all of these techniques. Results of all of these techniques when used separately, and combined is explained further.

4.5 Cosine Similarity Measure

Term vectors from each document are compared to term vectors of the sample document one-by-one on the basis of the following mathematical equation:

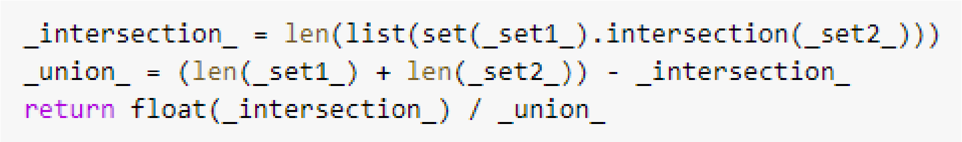


```
# cosine formula
for i in range(len(_union_)):
        c+= l1[i]*l2[i]
cosine = c / float((sum(l1)*sum(l2))**0.5)
return cosine
```

Figure 4.2: Cosine Similarity Calculation Python

4.6 Jaccard Similarity Measure

Term vectors from each document are compared to term vectors of the sample document one-by-one on the basis of the following mathematical equation:



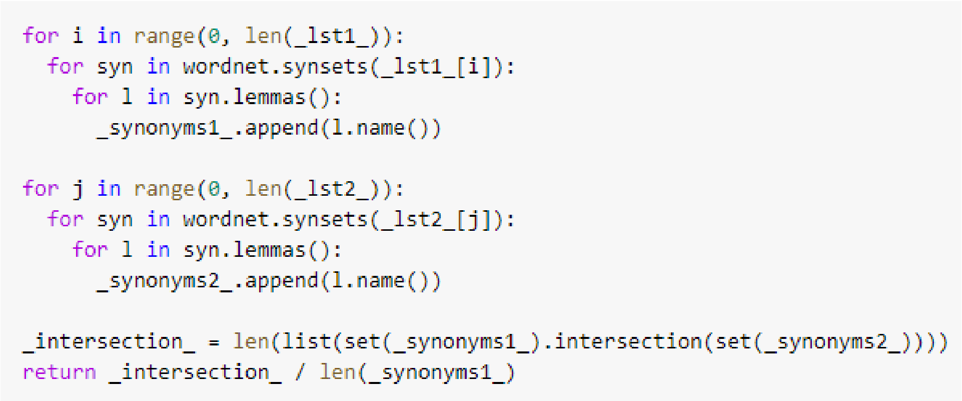
```
_intersection_ = len(list(set(_set1_).intersection(_set2_)))
_union_ = (len(_set1_) + len(_set2_)) - _intersection_
return float(_intersection_) / _union_
```

FIGURE 4.3: Jaccard Similarity Calculation Python

4.7 String Similarity Measure

The headings from each document are compared to the headings of sample document in two steps. This similarity score only focuses on the headings in order to test the documents on the basis of structure. In the first step, exact phrase matching is performed. In the second step, synonym matching is performed. The average score of both of the previous score is considered as string similarity score.

$$\text{Ssim} = \frac{\text{EPMScore} + \text{SMScore}}{2} \tag{4.1}$$



```
for i in range(0, len(_lst1_)):
  for syn in wordnet.synsets(_lst1_[i]):
    for l in syn.lemmas():
      _synonyms1_.append(l.name())

for j in range(0, len(_lst2_)):
  for syn in wordnet.synsets(_lst2_[j]):
    for l in syn.lemmas():
      _synonyms2_.append(l.name())

_intersection_ = len(list(set(_synonyms1_).intersection(set(_synonyms2_))))
return _intersection_ / len(_synonyms1_)
```

FIGURE 4.4: String Similarity Calculation with the help of wordnet Python

4.7.1 Weighted Sum based Similarity Calculation

The weighted sum of all above results is calculated as follows:

$$\begin{aligned} &Score(D1, D2) \\ &= (CosineSim(D1, D2) \times W1) + (JaccardSim(D1, D2) \times W2) \\ &+ (StringSim(D1, D2) \times W3) \end{aligned} \tag{4.2}$$

Table 4.1: Similarity Scores of all test documents w.r.t the standard template

Documents	CosineSim	JaccardSim	StringSim	Similarity Score
Test 1	0.29	0.17	0.30	0.27
Test 2	0.27	0.17	0.22	0.25
Test 3	0.29	0.17	0.22	0.22
Test 4	0.30	0.16	0.33	0.30
Test 5	0.24	0.15	0.16	0.18
Test 6	0.29	0.20	0.21	0.22
Test 7	0.34	0.25	0.31	0.30
Test 8	0.26	0.15	0.21	0.21
Test 9	0.23	0.15	0.19	0.19
Test 10	0.20	0.15	0.18	0.18

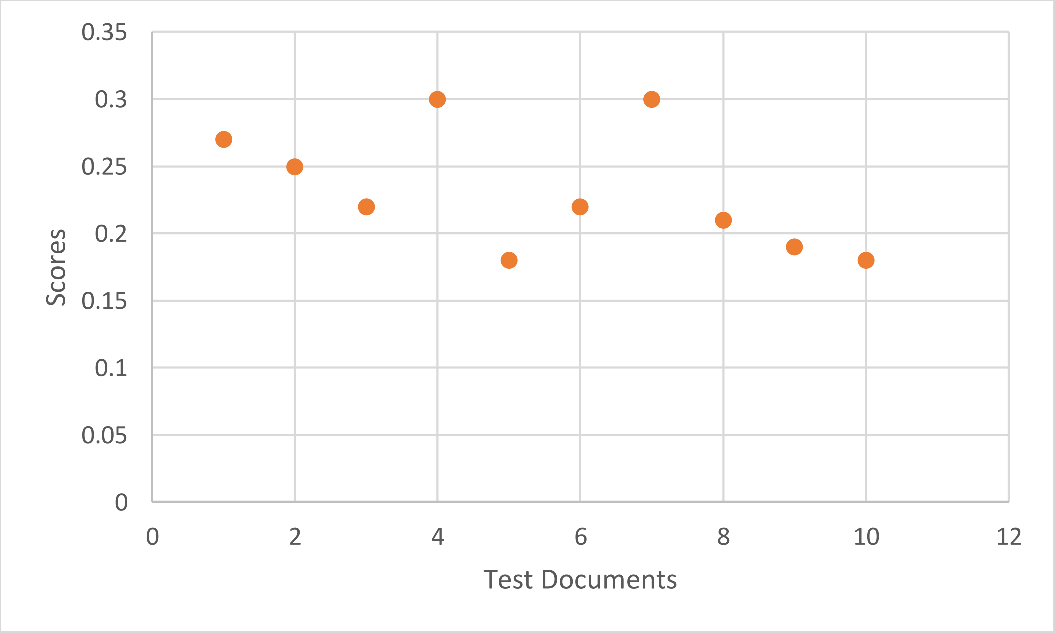


Figure 4.5: Scatter Graph of Similarity Scores of Proposed Technique

4.8 Results

According to Cosine Similarity Measure the document test 7 is most similar to the standard document while the document test 5 and test 10 are least similar. According to Jaccard Similarity Measure the document test 7 is most similar to the standard document while the documents test 5 and test 10 are least similar. According to String-based Similarity Measure the document test 7 is most similar to the standard document while the document test 5 is least similar. According to the proposed technique, the documents test 4 and test 7 are most similar to the standard document while the documents test 5 and test 10 are least similar. A cosine value of 0 means that the two documents are orthogonal and have no match. The closer the cosine value to 1, the smaller the angle and the greater the match between documents. Smaller values of Jaccard Similarity Metrics may indicate lesser similarity between documents but these values can also be erroneous as the sample was not very big. Higher values of String-based Similarity Measure indicate that there were more exact matches in the document. It can be observed that there is not very large difference between the values obtained from Cosine Similarity Measure and the values obtained from String-based Similarity Measure. However, the values obtained from the proposed technique match the most with those of Cosine Similarity Measure.

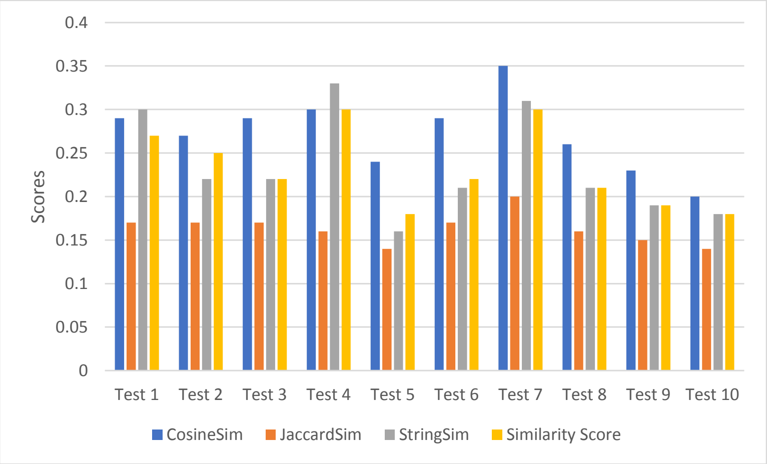


FIGURE 4.6: Comparison of Scores from each technique

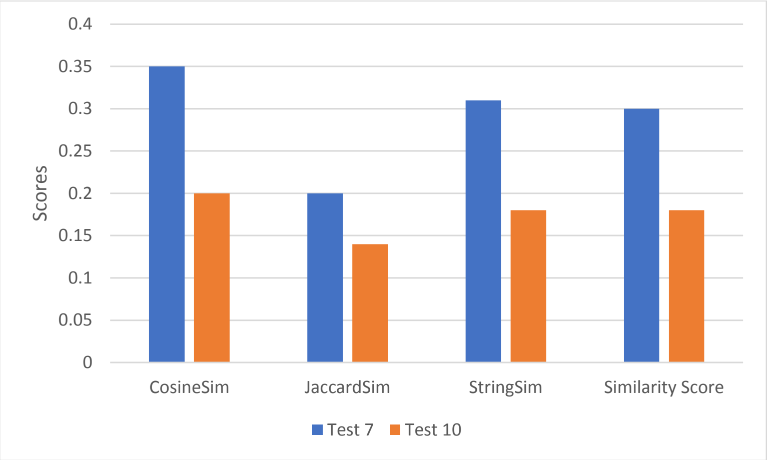


FIGURE 4.7: Least vs Most Similar test document

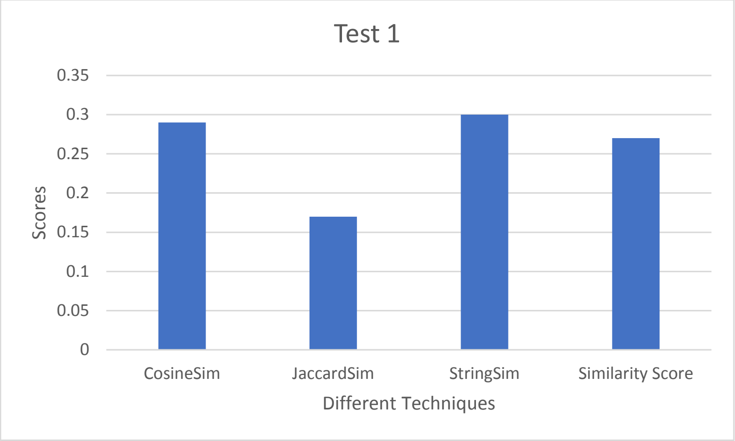


FIGURE 4.8: Comparison of Results from Different Techniques for Document Test 1

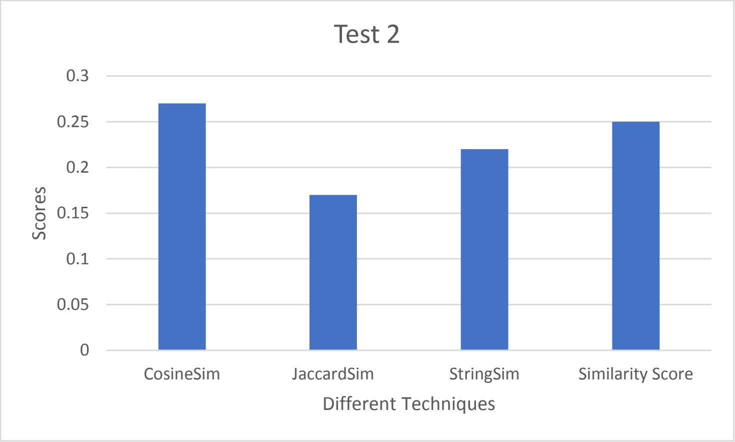


FIGURE 4.9: Comparison of Results from Different Techniques for Document Test 2

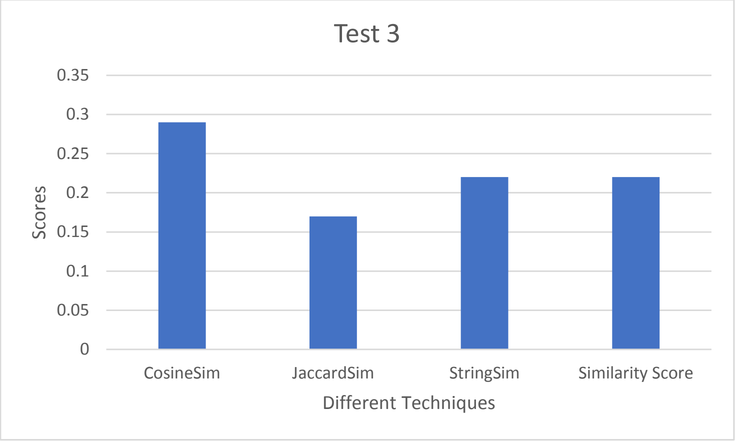


FIGURE 4.10: Comparison of Results from Different Techniques for Document Test 3

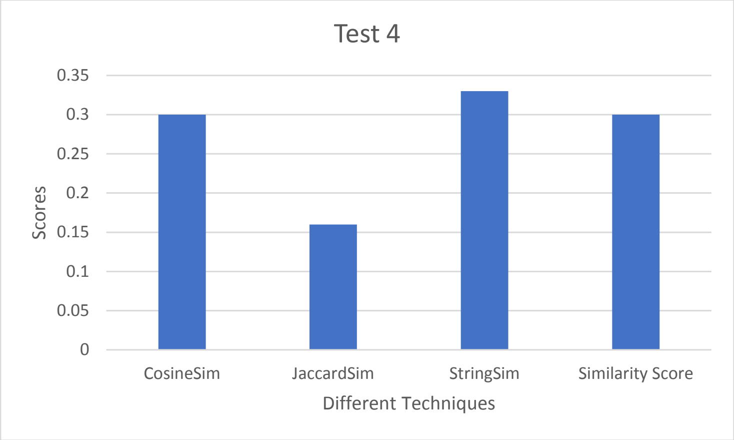


FIGURE 4.11: Comparison of Results from Different Techniques for Document Test 4

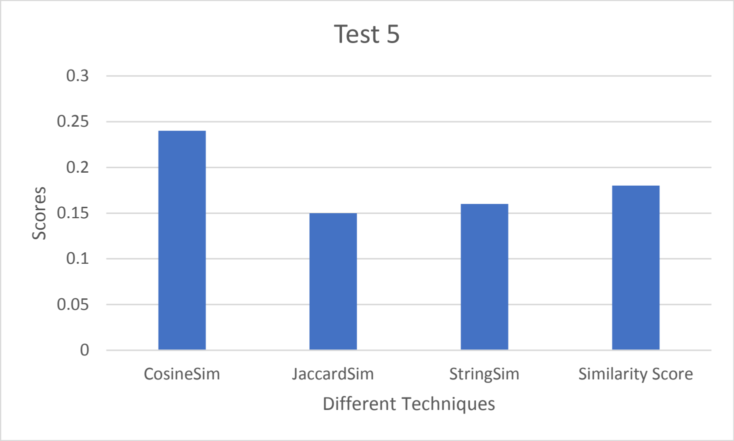


FIGURE 4.12: Comparison of Results from Different Techniques for Document Test 5

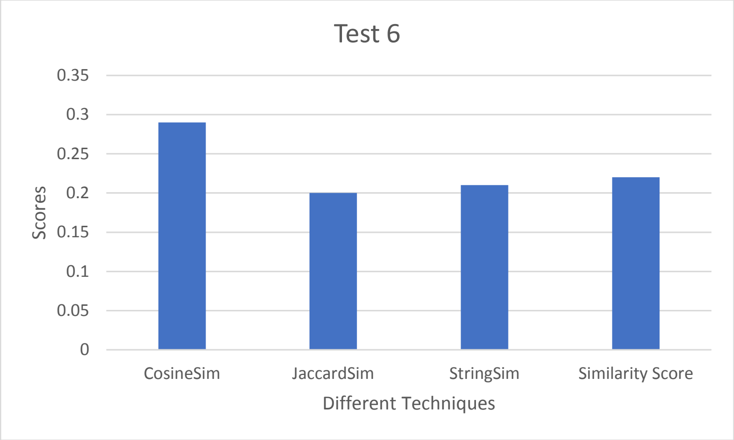


FIGURE 4.13: Comparison of Results from Different Techniques for Document Test 6

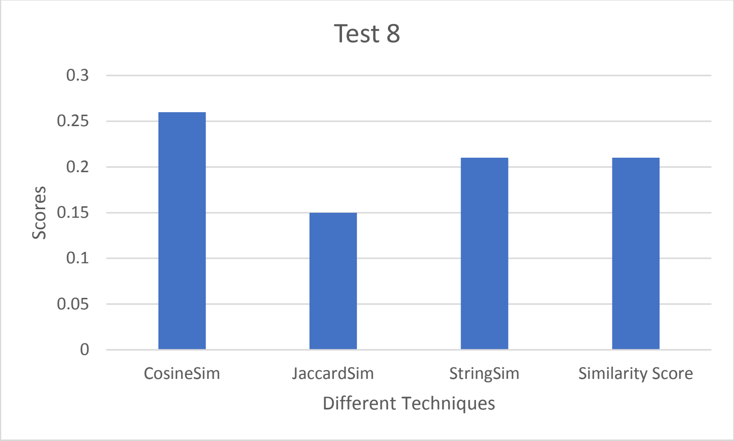


FIGURE 4.14: Comparison of Results from Different Techniques for Document Test 8

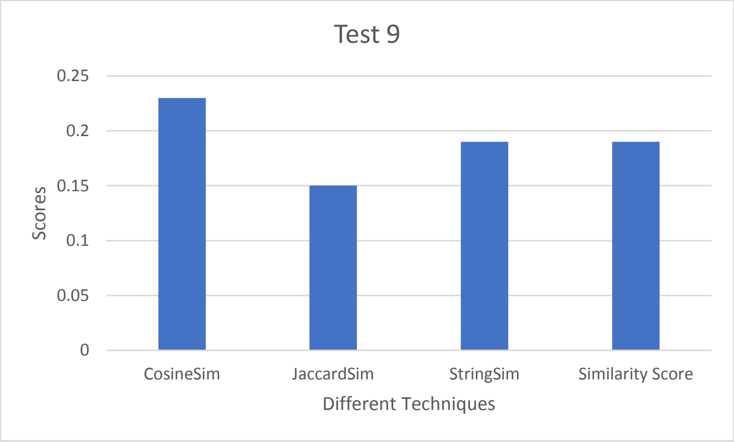


FIGURE 4.15: Comparison of Results from Different Techniques for Document Test 9

The results show that Cosine Similarity Measure, String Similarity Measure and the proposed technique compute results that are somewhat similar to each other and Jaccard Similarity Measure gives lesser scores as compared to all of them. The reason for this can be the nature of data that is being used in this experiment. Jaccard Similarity Measure does not work very well with nominal data. It is also known to give erroneous results with smaller samples. The values produced in proposed technique are more similar to string-based similarity calculation than that of cosine. This can be because of the weights that were assigned to each similarity measure as Sting-Based Similarity Measure was given the maximum weight.

4.9 Results Evaluation

To check the authenticity of this technique, the results obtained from the technique are evaluated in two different ways. For the first evaluation method, a user-based test is performed and its results are compared to the results of proposed technique. The second method used for the evaluation of technique is the use of standard evaluation measures, i.e., Accuracy, Precision, Recall and F-Measure.

4.9.1 User-Based Evaluation

Document Similarity Techniques and Compliance Identification are largely discussed in previous literature but there is no previous research work which identifies the compliance of information security documents using this technique. Therefore, user-based evaluation is being considered as gold standard for this research. The documents involved in this experiment are reviewed by 4 persons separately. 2 of these persons are students of information security while the other 2 do not have any affiliation with the field of information security. They were given a brief introduction about the topic and purpose of research. They were asked to score the documents for similarities in range of 0 to 1. Every person went through each document and produced a score for each document manually. The similarity scores evaluated by each person can be observed in the table below.

Table 4.2: User-study based Similarity Scores

Documents	Person 1	Person 2	Person 3	Person 4	Average
Test 1	0.30	0.25	0.33	0.30	0.29
Test 2	0.25	0.25	0.30	0.35	0.28
Test 3	0.20	0.30	0.25	0.28	0.24
Test 4	0.35	0.38	0.35	0.28	0.34
Test 5	0.20	0.35	0.25	0.40	0.27
Test 6	0.30	0.20	0.25	0.25	0.25
Test 7	0.35	0.35	0.35	0.30	0.34
Test 8	0.25	0.25	0.20	0.20	0.22
Test 9	0.30	0.20	0.35	0.25	0.28
Test 10	0.15	0.20	0.25	0.20	0.20

According to Person 1, the least similar documents is test 10 and the most similar document is test 4.

According to Person 2, the least similar documents are test 6, test 9 and test 10 and the most similar document is test 4.

According to Person 3, the least similar documents are test 3, test 5, test 6 and test 7 while the most similar document is test 7.

According to Person 4, the least similar documents are test 8 and test 10 and the most similar documents are test 5 and test 7.

According to the average of all results, test 10 is the least similar document, while test 10 is the most similar document. This result is very similar to the results obtained from the proposed technique.

The values are rounded off in order to calculate the best possible values of evaluation measures of this result. The scores from proposed technique and average of person-based evaluation are not exactly the same but they can be said as approximately same when rounding them off gives the same score.

Table 4.3: Comparison of Results of Proposed Technique with user-study based scores

Documents	Average of Scores by Persons	Proposed Technique Scores	Rounded Off Scores (Persons)	Rounded Off Scores (Technique)
Test 1	0.29	0.27	0.30	0.30
Test 2	0.28	0.25	0.30	0.30
Test 3	0.24	0.22	0.20	0.20
Test 4	0.34	0.30	0.30	0.30
Test 5	0.27	0.18	0.30	0.20
Test 6	0.25	0.22	0.30	0.20
Test 7	0.34	0.30	0.30	0.30
Test 8	0.22	0.21	0.20	0.20
Test 9	0.28	0.19	0.30	0.20
Test 10	0.20	0.18	0.20	0.20

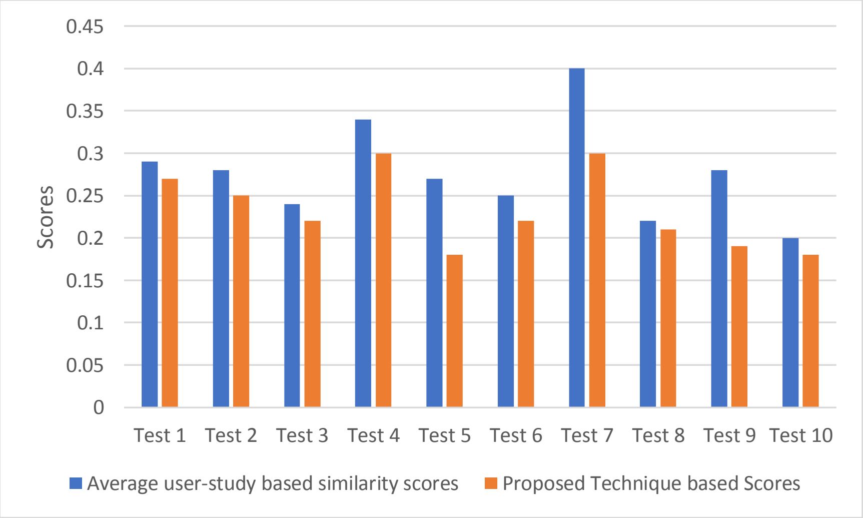


Figure 4.16: Evaluation of Results of proposed technique

Small differences are observed in the similarity scores of documents test 1, test 2, test 3, test 6, test 8 and test 10, while test 4 shows a slight change and test 5, test 7 and test 9 show a greater difference. The values are rounded off in order to calculate the evaluation measures of this result.

Table 4.4: Comparison of Results of Cosine Similarity Measure with user-study based scores

Documents	Average of Scores by Persons	Proposed Technique Scores	Rounded Off Scores (Persons)	Rounded Off Scores (Technique)
Test 1	0.29	0.29	0.30	0.30
Test 2	0.28	0.27	0.30	0.30
Test 3	0.24	0.29	0.20	0.30
Test 4	0.34	0.30	0.30	0.30
Test 5	0.27	0.24	0.30	0.20
Test 6	0.25	0.29	0.30	0.30
Test 7	0.34	0.34	0.30	0.30
Test 8	0.22	0.26	0.20	0.30
Test 9	0.28	0.23	0.30	0.20
Test 10	0.20	0.20	0.20	0.20

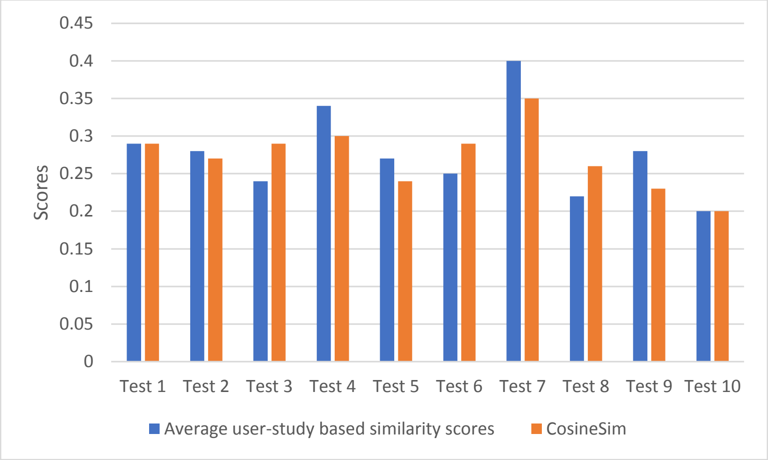


FIGURE 4.17: Evaluation of Results of Cosine Similarity

The scores of documents test 1 and test 10 obtained from Cosine Similarity technique and that of user-based evaluation are the same. Small differences are observed in scores of documents test 2 and test 5 while the documents test 3, test 4, test 6, test 8 and test 9 show slightly greater differences and test 7 shows maximum difference.

TABLE 4.5: Comparison of Results of Jaccard Similarity Measure with user-study based scores

Documents	Average user-study based similarity scores	Jaccard Similarity Scores	Rounded Off Scores (Persons)	Rounded Off Scores (Technique)
Test 1	0.29	0.17	0.30	0.20
Test 2	0.28	0.17	0.30	0.20
Test 3	0.24	0.17	0.20	0.20
Test 4	0.34	0.16	0.30	0.20
Test 5	0.27	0.15	0.30	0.20
Test 6	0.25	0.20	0.30	0.20
Test 7	0.34	0.25	0.30	0.30
Test 8	0.22	0.16	0.20	0.20
Test 9	0.28	0.15	0.30	0.20
Test 10	0.20	0.15	0.20	0.20

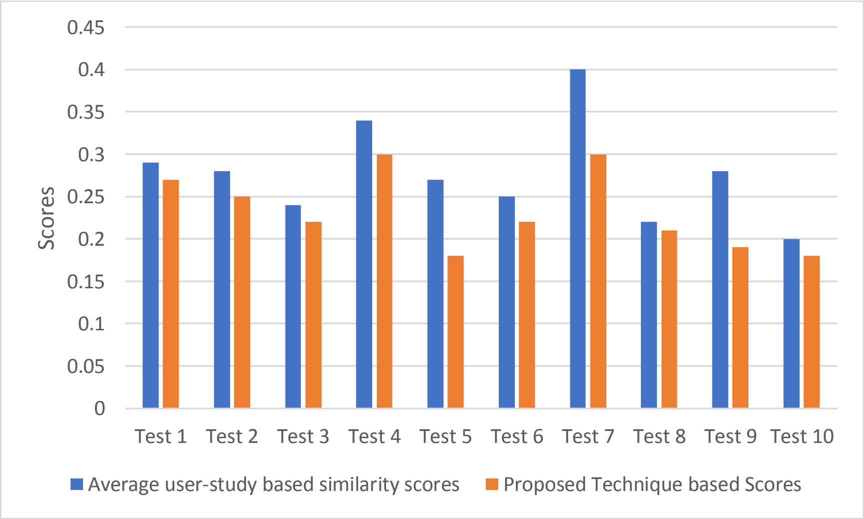


FIGURE 4.18: Evaluation of Results of Jaccard Similarity

Large differences are observed between the results obtained from Jaccard Similarity technique and that of user-based evaluation. These results dont match with the results obtained from any other technique. This means that Jaccard Similarity coefficient alone can not compute similarity between documents specially when the sample is small.

TABLE 4.6: Comparison of Results of String-based Measure with user-study based scores

Documents	Average of Scores by Persons	Proposed Technique Scores	Rounded Off Scores (Persons)	Rounded Off Scores (Technique)
Test 1	0.29	0.30	0.30	0.30
Test 2	0.28	0.22	0.30	0.20
Test 3	0.24	0.22	0.20	0.20
Test 4	0.34	0.33	0.30	0.30
Test 5	0.27	0.16	0.30	0.20
Test 6	0.25	0.21	0.30	0.20
Test 7	0.34	0.31	0.30	0.30
Test 8	0.22	0.21	0.20	0.20
Test 9	0.28	0.19	0.30	0.20
Test 10	0.20	0.18	0.20	0.20

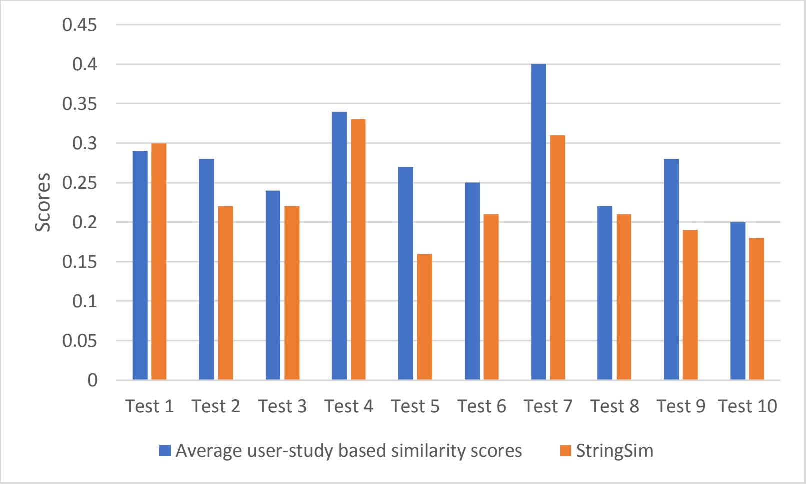


FIGURE 4.19: Evaluation Results of String-Based Similarity

The scores of documents test 1, test 3, test 4, test 8 and test 10 show very minor differences while that of documents test 2, test 5, test 6, test 7 and test 9 show greater differences while test 7 and test 9 being the most deviant.

Out of all 4 similarity techniques, the evaluated scores match the most with the proposed technique and the next most similar scores are of cosine similarity measure. However, String-based similarity is still showing better results than Jaccard similarity measure. For the calculation of evaluation measures, the score 0.3 is taken as a positive and 0.2 is taken as a negative as all of the results are in the range of 0.2-0.3. According to this assumption, the cases where both evaluated and proposed results are 0.3 are said to be true positive and the cases where both evaluated and proposed results are 0.2 are said to be true negative. Moreover, the cases where evaluated scores are in the range of 0.2 but results from the proposed technique are in the range of 0.3 are said to be false positive and the cases where evaluated scores are in the range of 0.3 but results from the proposed technique are in the range of 0.2 are said to be false negative.

4.9.2 Standard Evaluation Measures

The measures discussed in the following text are commonly used in literature to evaluate the effectiveness of any proposed model or technique.

4.9.2.1 Confusion Matrix

A confusion matrix is a summary of the outcomes of prediction over an issue of classification.

The number of correct and incorrect predictions was summarized and broken down by each class by counting values. This is the key to the matrix of confusion.

True Positive (TP): Number of positive samples correctly labeled
True Negative (TN): Number of Negative Samples correctly labeled
False Positive (FP): Number of negative samples incorrectly labelled as positive
False Negative (FN): Number of positive samples incorrectly labelled as negative

Table 4.7: Confusion Matrix for Proposed Technique

Predicted vs Actual Values	Positive	Negative
Positive	4	3
Negative	0	3

The confusion matrix of proposed technique shows that out of 10 results, 4 of them are true positives, 3 of them are true negative and 3 of them are false positive.

Table 4.8: Confusion Matrix for Cosine Similarity Measure

Predicted vs Actual Values	Positive	Negative
Positive	5	2
Negative	2	1

The confusion matrix of cosine similarity technique shows that:

Out of 10 results, 5 of them are true positives, One of them is true negative and two of them are false positive and two of them are false negative.

Table 4.9: Confusion Matrix for Jaccard Similarity Measure

Predicted vs Actual Values	Positive	Negative
Positive	1	0
Negative	6	3

The confusion matrix of Jaccard similarity technique shows that:

Out of 10 results, 1 of them is true positive,

6 of them are false negatives and 3 of them true negative.

No false positive value is found.

Table 4.10: Confusion Matrix for String-based Similarity Measure

Predicted vs Actual Values	Positive	Negative
Positive	3	4
Negative	0	3

The confusion matrix of string-based similarity technique shows that:

Out of 10 results, 3 of them are true positives,

3 of them are true negative,

4 of them are false positive and no false negative value is found.

4.9.2.2 Accuracy

Accuracy is the measure of correctly predicted scores.

It is calculated with the help of the following formula:

$$
\begin{aligned}
Accuracy &= \frac{Correct\ Prediction\ Count}{Total\ number\ of\ Preditions} \\
&TheAccuracyofProposedTechnique: \\
Accuracy &= \frac{7}{10} = 0.7 \\
&TheAccuracyofCosineSimilarityTechnique: \\
Accuracy &= \frac{6}{10} = 0.6 \\
&TheAccuracyofJaccardSimilarityTechnique: \\
Accuracy &= \frac{4}{10} = 0.4 \\
&TheAccuracyofString-basedSimilarityTechnique: \\
Accuracy &= \frac{6}{10} = 0.6
\end{aligned}
\tag{4.3}
$$

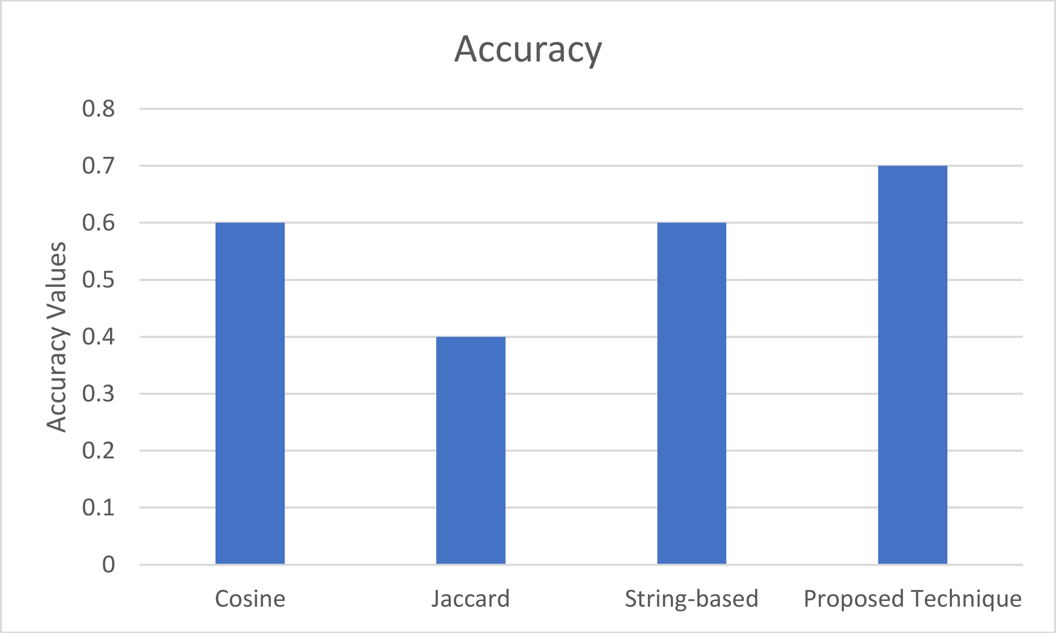


Figure 4.20: Comparison of Accuracy values of all Techniques

4.9.2.3 Precision

Precision is the measure of correct positive predictions. It is calculated with the help of the following formula:

$$
\begin{aligned}
&Precision \;=\; \frac{\sum TP}{\sum TP \;+\; \sum FP}\\
&Where TP = True Positives and FP = False Positives\\
&The Precision of proposed technique:\\
&Precision = \frac{4}{7} = \; 0.6\\
&The Precision of Cosine Similarity technique:\\
&Precision = \frac{5}{7} = \; 0.7\\
&The Precision of Jaccard Similarity technique:\\
&Precision = \frac{1}{1} = \; 1\\
&The Precision of String - based Similarity technique:\\
&Precision = \frac{3}{7} = \; 0.4
\end{aligned}
\tag{4.4}
$$

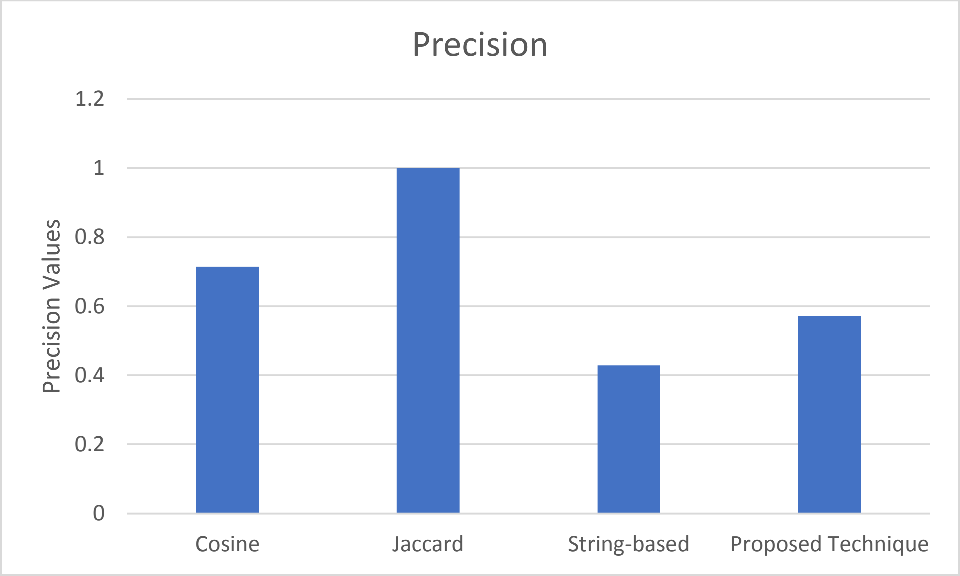


FIGURE 4.21: Comparison of Precision values of all Techniques

4.9.2.4 Recall

Recall is used to measure the extent of actual positives that are identified correctly. It is calculated with the help of the following formula:

$$\begin{aligned}
&Recall \;=\; \frac{\sum TP}{\sum TP \;+\; \sum FN} \\
&Where TP = True Positives and FN = False Negative \\
&The Recall value of proposed technique : \\
&Recall = \frac{4}{4} = \;1 \\
&The Recall value of Cosine Similarity technique : \\
&Recall = \frac{5}{7} = \;0.7 \\
&The Recall value of Jaccard Similarity technique : \\
&Recall = \frac{1}{7} = \;0.1 \\
&The Recall value of String-based Similarity technique : \\
&Recall = \frac{4}{4} = \;1
\end{aligned} \tag{4.5}$$

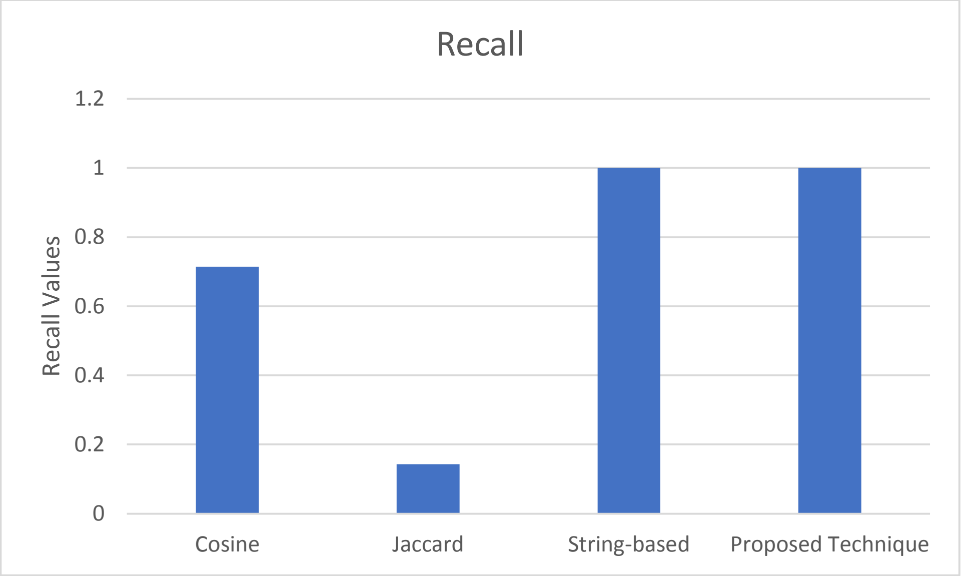


Figure 4.22: Comparison of Recall values of all Techniques

4.9.2.5 F Measure

F-Measure combines both precision and recall into a single measure that comprises of both properties. It is calculated with the help of the following formula:

$$
\begin{aligned}
&\mathrm{F-Measure} \ = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \\
&TheF - Measure value of proposed technique : \\
&F - Measure = \frac{(2 \times 0.6 \times 1)}{0.6 + 1} = \ 0.73 \\
&TheF - Measure value of Cosine Similarity Technique : \\
&F - Measure = \frac{(2 \times 0.7 \times 0.7)}{0.7 + 0.7} = \ 0.70 \\
&TheF - Measure value of Jaccard Similarity Technique : \\
&F - Measure = \frac{(2 \times 1 \times 0.1)}{1 + 0.1} = \ 0.25 \\
&TheF - Measure value of String - based Similarity Technique : \\
&F - Measure = \frac{(2 \times 0.4 \times 1)}{0.4 + 1} = \ 0.60
\end{aligned}
\tag{4.6}
$$

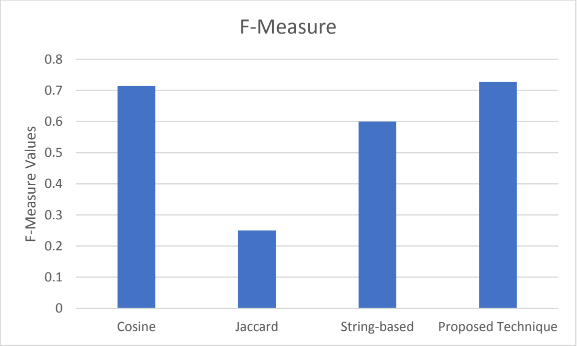


Figure 4.23: Comparison of F-Measure values of all Techniques

The above calculations are concluded in the following table:

Table 4.11: Comparison of results from Evaluation Standards

	Cosine Similarity Technique	Jaccard Similarity Technique	String Similarity Technique	Proposed Technique
Accuracy	0.6	0.4	0.6	0.7
Precision	0.7	1	0.4	0.6
Recall	0.7	0.1	1	1
F-Measure	0.70	0.25	0.60	0.73

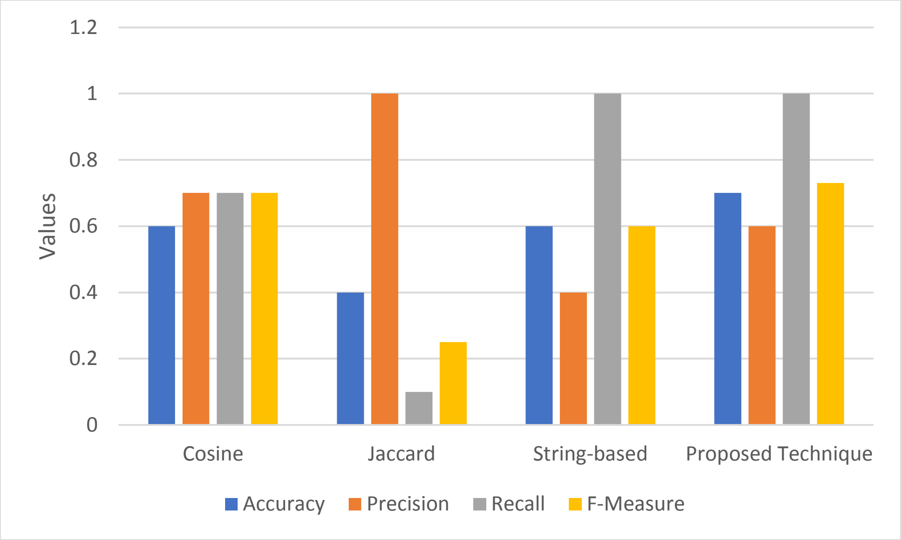


FIGURE 4.24: Overall Comparison of all Results

The proposed technique has the highest value of Accuracy whereas the Jaccard Similarity technique has the lowest value of Accuracy. Although the value of Accuracy proves the proposed technique to be most efficient, it is important to take other evaluation measures in account too because accuracy alone cannot guarantee the validity of a technique specially when the results are not symmetric, i.e., the number of true positives are not equal to number of true negatives.

The precision score of the Cosine similarity technique is very similar to that of proposed technique. Precision value 1 indicates that all of the positive samples are classified as positive samples and none of the positive samples are classified incorrectly. The Precision of Jaccard Similarity technique is found to be 0 because no true positive or true negative cases were detected.

The value of recall of String-based technique and the proposed technique are found to be maximum. The recall score of 1.0 means that all relevant information was retrieved along with the irrelevant information. Recall score only depends on the extent of relative information found. As there was no False Negative case detected in case of String-based technique and the proposed technique, the recall score is computed to be exactly 1. The Recall value of Cosine Similarity technique is good but lesser than that of proposed technique similarity. The Recall value of Jaccard Similarity technique is found to be minimum.

F-Measure is a combined metric and its value depends upon the values of Precision and Recall. Since the Precision score of Cosine Similarity technique is maximum and the Recall score of Proposed technique is maximum, hence their F-Measure score is also maximum.

4.10 Discussions

- The test documents test 5 and test 10 are found to be least similar to the standard template in the results of proposed technique as well as the results of user-based evaluation. Similarly, test document test 7 is found to be the most similar in every approach.

- The accuracy of model is calculated to be 0.7. In simple words, it means that the technique gives 70% accurate results. But Accuracy alone is not enough to check validity of a technique.

- The 0.6 score of precision is interpreting that 60% of its results computed by the technique are relevant, i.e., 60% precise results are computed.

- The value of recall is 1. It means that 100% relevant results were retrieved by the technique. However, it does not guarantee that all of the obtained results are relevant.

- Lastly, the F-Measure is calculated to be 0.73. This value is close to the value of Accuracy.

- The accuracy and recall scores of the proposed technique are better than the all-other techniques. Due to good recall score, the F-Measure score of the Proposed technique is also found to be maximum.

Chapter 5

Conclusions and Future Tasks

It is becoming more and more critical for organizations with every passing day to secure their constantly growing data properly. Information Security Policies help organizations to follow certain security procedures. These policies are written in ISP documents. Reviewing the whole ISP document and identifying its compliance to a security standard manually is a very hefty task. This problem of Non-compliance of information security policy documents with standards of security is addressed in this study. A detailed literature review is performed in order to find more about previously being used techniques and identify the significance of the discussed problem. A new technique is proposed to identify the compliance of ISPD with a standard document.

The research work can be concluded as:

1. The most common techniques found in literature used for primary data analysis for the similarity calculation are Cosine Similarity Measure, Jaccard Similarity Measure, etc. Each technique works efficiently in different cases to extract diverse information from the data. Retrieval of the most similar texts to a given document generally function better with cosine similarity, while Jaccard similarity is good for cases where duplication does not matter. Hardly any technique was found to identify compliance of ISP Documents in the prior literature.

2. However, the new proposed technique made use of Cosine Similarity Measure, Jaccard Similarity Measure and String Similarity Measure. Weighted sum of all these 3 similarity measures computes as the final score. This technique can be used to improve the overall security of the organization because the first step towards security is the development of a sustainable security policy and its implementation.

3. 3. The new proposed technique computes better results as compared to the results of individual techniques. The evaluation of results is performed with the help of standard evaluation measures. The results from user-study based evaluation are considered to be gold-standard. The user-study is performed by 4 people with different backgrounds. The proposed model has the accuracy score of 0.7, precision 0.6, recall 1 and F-measure of 0.73.

4. It can be concluded that the overall accuracy of the similarity score can be increased by the combination of several similarity measures instead of using them separately. Moreover, the proposed technique provides a fast way to compute similarity scores of any given number of documents as the technique used is scalable.

5. Information Security policy developers will benefit from this work. Organizations will be able to check the legitimacy of their security policy documents very easily by using this technique. Generally, this study will help to improve security of organizations so it can be considered an addition to the field of information security technology.

Following are some of the potential directions for future research in this area that are identified:

1. Three specific techniques are used in this study. Different combinations of techniques can be applied to it in future to explore more information from the security policy documents and identify which combination of technique works best with ISPDs.

2. This technique is identifying the extent of similarity between two documents in the form of a score in range 0 to 1. More functionalities can be applied in it to identify the exact lines where a flaw is present.

3. This experiment is performed on the documents from the domain of health sector. Documents from various other domains can be tested by using this technique.

4. A greater number of documents can be used to perform the same experiment to further improve the results.

Bibliography

[1] M. P. Buthelezi, J. A. Van Der Poll, and E. O. Ochola, "Ambiguity as a barrier to information security policy compliance: A content analysis," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI).* IEEE, 2016, pp. 1360–1367.

[2] V. Chenthamarakshan, R. A. Hosn, S. Ikbal, N. Kambhatla, D. Majumdar, and S. Sarkar, "Measuring compliance and deviations in a template-based service contract development process," in *2010 IEEE International Conference on Services Computing.* IEEE, 2010, pp. 289–296.

[3] A. Sayeed, S. Sarkar, Y. Deng, R. Hosn, R. Mahindru, and N. Rajamani, "Characteristics of document similarity measures for compliance analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1207–1216.

[4] S. Hina and P. D. D. Dominic, "Information security policies compliance: a perspective for higher education institutions," *Journal of Computer Information Systems*, 2018.

[5] A. Bhardwaj, G. Subrahmanyam, V. Avasthi, and H. Sastry, "Design a resilient network infrastructure security policy framework," *Indian Journal of Science and Technology*, vol. 9, no. 19, pp. 1–8, 2016.

[6] K. M. Hammouda and M. S. Kamel, "Phrase-based document similarity based on an index graph model," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 2002, pp. 203–210.

[7] E. Rostami, F. Karlsson, and E. Kolkowska, “The hunt for computerized support in information security policy management,” *Information & Computer Security*, 2020.

[8] M. Siponen, “Six design theories for is security policies and guidelines,” *Journal of the Association for Information systems*, vol. 7, no. 1, p. 19, 2006.

[9] E. Rostami, F. Karlsson, and S. Gao, “Requirements for computerized tools to design information security policies,” *Computers & Security*, vol. 99, p. 102063, 2020.

[10] J. Coertze and R. von Solms, “A software gateway to affordable and effective information security governance in smmes,” in *2013 Information Security for South Africa*. IEEE, 2013, pp. 1–8.

[11] C. Vermeulen and R. Von Solms, “The information security management toolbox–taking the pain out of security management,” *Information management & computer security*, 2002.

[12] G. Liu and H. Zhang, “An ontology constructing technology oriented on massive social security policy documents,” *Cognitive Systems Research*, vol. 60, pp. 97–105, 2020.

[13] R. A. Alias *et al.*, “Information security policy compliance: Systematic literature review,” *Procedia Computer Science*, vol. 161, pp. 1216–1224, 2019.

[14] I. Meriah and L. B. A. Rabai, “Comparative study of ontologies based iso 27000 series security standards,” *Procedia Computer Science*, vol. 160, pp. 85–92, 2019.

[15] M. Park and S. Chai, “Internalization of information security policy and information security practice: A comparison with compliance,” in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[16] S. Patel, A. Makwana, S. Jardosh, and I. C. Changa, “Analysis and survey on string matching algorithms for ontology matching.”

[17] M. Kang, T. Lee, and S. Um, "Establishment of methods for information security system policy using benchmarking," in *2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 2018, pp. 237–242.

[18] J. Amsenga, "An introduction to standards related to information security." in *ISSA*, 2008, pp. 1–18.

[19] H. Paananen, M. Lapke, and M. Siponen, "State of the art in information security policy development," *Computers & Security*, vol. 88, p. 101608, 2020.

[20] K. Höne and J. Eloff, "What makes an effective information security policy?" *Network security*, vol. 2002, no. 6, pp. 14–16, 2002.

[21] D. Danchev, "Building and implementing a successful information security policy," *online at www. windowsecurity. com*, 2003.

[22] K. Höne and J. H. P. Eloff, "Information security policywhat do international information security standards say?" *Computers & security*, vol. 21, no. 5, pp. 402–409, 2002.

[23] J. Järveläinen, "Integrated business continuity planning and information security policy development approach," 2016.

[24] K. J. Knapp, R. F. Morris Jr, T. E. Marshall, and T. A. Byrd, "Information security policy: An organizational-level process model," *Computers & security*, vol. 28, no. 7, pp. 493–508, 2009.

[25] R. Diesch, M. Pfaff, and H. Krcmar, "A comprehensive model of information security factors for decision-makers," *Computers & Security*, vol. 92, p. 101747, 2020.

[26] M. Montanari, E. Chan, K. Larson, W. Yoo, and R. H. Campbell, "Distributed security policy conformance," in *IFIP International Information Security Conference*. Springer, 2011, pp. 210–222.

[27] K. Kurtel, "Information security policy: positioning the technological components of information security services under the perspective of electronic business," in *Security of Information and Networks: Proceedings of the First International Conference on Security of Information and Networks (SIN 2007).* Trafford Publishing, 2008, p. 302.

[28] M. S. Ofori-Duodu, "Exploring data security management strategies for preventing data breaches," 2019.

[29] T. Tuyikeze and D. Pottas, "An information security policy development life cycle," in *Proceedings of the South African Information Security Multi-Conference (SAISMC), Port Elizabeth, South Africa*, 2011, pp. 165–176.

[30] A. Hovav *et al.*, "Benchmarking methodology for information security policy (bmisp): Artifact development and evaluation," *Information Systems Frontiers*, vol. 22, no. 1, pp. 221–242, 2020.

[31] V. Gowadia, C. Farkas, and M. Kudo, "Checking security policy compliance," *arXiv preprint arXiv:0809.5266*, 2008.

[32] D. J. Simms, "Information security optimization: from theory to practice," in *2009 International Conference on Availability, Reliability and Security.* IEEE, 2009, pp. 675–680.

[33] S. Talbot and A. Woodward, "Improving an organisations existing information technology policy to increase security," 2009.

[34] R.-Y. Ye and L.-J. Feng, "Technical and economic models of information security," in *2015 International Conference on Computer Science and Applications (CSA).* IEEE, 2015, pp. 329–332.

[35] K. Arbanas and N. Žajdela Hrustek, "Key success factors of information systems security," *Journal of Information and Organizational Sciences*, vol. 43, no. 2, pp. 131–144, 2019.

[36] J. O. Oyelami and A. M. Kassim, "Cyber security defence policies: A proposed guidelines for organisations cyber security practices," *International*

Journal of Advanced Computer Science and Applications, vol. 11, no. 8, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110817

[37] A. Singh, C. Ramakrishnan, I. Ramakrishnan, S. D. Stoller, and D. S. Warren, "Security policy analysis using deductive spreadsheets," in *Proceedings of the 2007 ACM workshop on Formal methods in security engineering*, 2007, pp. 42–50.

[38] H. Liu and P. Wang, "Assessing text semantic similarity using ontology." *JSW*, vol. 9, no. 2, pp. 490–497, 2014.

[39] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 19–28, 2016.

[40] G. Salton, "Automatic text processing: The transformation, analysis, and retrieval of," *Reading: Addison-Wesley*, vol. 169, 1989.

[41] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi, "The similarity metric," *IEEE transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[42] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *International conference on intelligent data engineering and automated learning*. Springer, 2013, pp. 611–618.

[43] L. Zahrotun, "Comparison jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method," *Computer Engineering and Applications Journal*, vol. 5, no. 1, pp. 11–18, 2016.

[44] S. Logeswari and K. Premalatha, "Biomedical document clustering using ontology based concept weight," in *2013 International Conference on Computer Communication and Informatics*, 2013, pp. 1–4.

[45] J. Mustafa, S. Khan, and K. Latif, "Ontology based semantic information retrieval," in *2008 4th International IEEE Conference Intelligent Systems*, vol. 3, 2008, pp. 22–14–22–19.

[46] W. H. Gomaa, A. A. Fahmy *et al.*, “A survey of text similarity approaches,” *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[47] M. Yu, G. Li, D. Deng, and J. Feng, “String similarity search and join: a survey,” *Frontiers of Computer Science*, vol. 10, no. 3, pp. 399–417, 2016.

[48] A. Madylova and S. G. Oguducu, “A taxonomy based semantic similarity of documents using the cosine measure,” in *2009 24th International Symposium on Computer and Information Sciences.* IEEE, 2009, pp. 129–134.

[49] S. Sohangir and D. Wang, “Improved sqrt-cosine similarity measurement,” *Journal of Big Data*, vol. 4, no. 1, pp. 1–13, 2017.

[50] A. K. Patidar, J. Agrawal, and N. Mishra, “Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach,” *International Journal of Computer Applications*, vol. 40, no. 16, pp. 1–5, 2012.

[51] T. Tuyikeze and S. Flowerday, “Information security policy development and implementation: A content analysis approach.” in *HAISA*, 2014, pp. 11–20.