“User and infrastructure security and privacy with regard to compliance”

A dissertation submitted to the Department of Digital Systems, of the University of Piraeus,

in fulfillment of the requirements for the degree of

Doctor of Philosophy

by **Konstantinos Vavousis**

M.Sc. in Digital Systems Security

School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Piraeus, 2021

Advisory Committee

Christos Xenakis – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus (supervisor)

Konstantinos Lambrinoudakis – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Sokratis Katsikas – Professor, Department of Information Security and Communication Technology, Norwegian University of Science and Technology

Examination Committee

Christos Xenakis – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus (supervisor)

Konstantinos Lambrinoudakis – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Sokratis Katsikas – Professor, Department of Information Security and Communication Technology, Norwegian University of Science and Technology

Stefanos Gritzalis – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Michael Sfakianakis – Professor, Applied Informatics in Business Administration, Department of Business Administration, University of Piraeus

Nikitas-Marinos Sgouros – Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Dimosthenis Kyriazis – Associate Professor, School of Information and Communication Technologies, Department of Digital Systems, University of Piraeus

Acknowledgments

My journey at the University of Piraeus has been a fruitful and exciting phase of my life. I am grateful to my advisor, Professor Christos Xenakis, who has been supportive and inspirational all these years, since my master's degree. Christos always has in mind the best interests of his students and is a constant source of encouragement and motivation.

I have also had an amazing experience working with Professor Christoforos Dadoyan, throughout my Ph.D. It was a truly enjoyable experience interacting with Christoforos and the Digital Systems lab. I am thankful to Dr. Marinos Papadopoulos, who introduced me to the world of security from the legal perspective. I thank my thesis committee: Christos Xenakis, Konstantinos Labrinoudakis, and Socratis Katsikas, from whom I have learned a great deal.

I would like to thank my amazing colleague Eleni Veroni, with whom I had the chance to collaborate on the worldwide password analysis for public Wi-Fi infrastructures and, hopefully, for more researches to come.

I would also like to thank my parents for all their support and patience since the day I was born and my brother who has always been there for me.

Finally, I am immeasurably indebted to my wife Andriani. She has always supported me unconditionally and has motivated me. This thesis is dedicated to her and our incredible daughters Anna and Dimitra.

Ευχαριστίες

Το ταξίδι μου στο Πανεπιστήμιο Πειραιά ήταν μια γόνιμη και συναρπαστική φάση της ζωής μου. Είμαι ευγνώμων στον επιβλέποντα, καθηγητή Χρήστο Ξενάκη, ο οποίος υπήρξε υποστηρικτικός όλα αυτά τα χρόνια, αρχής γενομένης από το μεταπτυχιακό μου. Ο κ. Ξενάκης γνωρίζει πως να ενθαρρύνει και να εμπνέει τους φοιτητές του και αποτελεί παράδειγμα προς μίμηση.

Είμαι επίσης ευγνώμων στην καταπληκτική συνεργασία με τον καθηγητή Χριστόφορο Νταντογιάν, καθ' όλη τη διάρκεια του διδακτορικού μου. Συνεργασία η οποία ήταν καθόλα εποικοδομητική όπως επίσης και συνολικά με τα μέλη του εργαστηρίου Digital Systems της Σχολής. Είμαι ευγνώμων στον Δρ. Μαρίνο Παπαδόπουλο, που με εισήγαγε στον κόσμο της ασφάλειας από την νομική σκοπιά. Θα ήθελα να ευχαριστήσω επίσης τα μέλη της τριμελούς επιτροπής: Χρήστο Ξενάκη, Κωνσταντίνο Λαμπρινουδάκη και Σωκράτη Κατσίκα, από τους οποίους έχω μάθει πολλά.

Ευχαριστώ επίσης την καταπληκτική συνάδελφό μου Ελένη Βερώνη, με την οποία είχα την ευκαιρία να συνεργαστώ στην έρευνα A Worldwide Empirical Analysis of Wi-Fi Passwords.

Θα ήθελα επίσης να ευχαριστήσω τους γονείς μου για όλη την υποστήριξη και την υπομονή τους από την ημέρα που γεννήθηκα, καθώς και τον αδερφό μου που ήταν πάντα εκεί για μένα.

Τέλος, είμαι απεριόριστα ευγνώμων στη σύζυγό μου Ανδριανή, η οποία πάντα με προτρέπει και με παρακινεί άνευ όρων. Αυτή η διατριβή είναι αφιερωμένη στη σύζυγο μου και στις φανταστικές κόρες μας Άννα και Δήμητρα.

Abstract

Nowadays, millions of companies and billions of users worldwide rely on networks either wireless or wired for their daily work and entertainment. Due to the lack of privacy-by-design and the absence of strong security mechanisms, there are multiple ways for malicious users to penetrate networks and systems. Ubiquitous Networking and Global Internet, which has become more portable and accessible than ever before through private and publicly available IT infrastructures, make unauthorized access more feasible. This also generates serious security and privacy concerns due to a number of ensuing cyber threats, especially in case of Internet access via public Wi-Fi networks. In the described context, Internet security should and can play an important role towards protecting our everyday lives and online interactions. Yet, most users are unaware of these threats and the extent to which their privacy might be compromised.

Regulations, such as the General Data Protection Regulation (GDPR), have been established to safeguard and improve the privacy and security of users and IT infrastructures, enforcing the installation of adequate cybersecurity measures. The application of regulations such as the GDPR is considered an issue of vital importance protecting the privacy and ensuring the security of IT infrastructures and websites, of data controllers and processors, both inside and outside the European Union. Such regulations may act as a useful toolset, which, among other requirements, mandates the adoption of privacy (and security)-by-design. While the GDPR implies a minimum set of technical Internet Security means to be taken into consideration by companies and organizations to achieve compliance, it is of high importance to highlight the adaptation of strong security mechanisms that will not only set companies compliant with the GDPR but also maintain them strong and resilient against multiple cyber threats.

In the present thesis, a big set of issues on privacy and security are analyzed, offering solutions to the numerous problems that companies and single users face either at work or in a recreational setting on a day to day basis. As a case study, in-depth IT security and privacy concerns regarding the National Library of Greece and the Greek Libraries Network of the National Library of Greece were examined, based on international

regulatory and IT security standards. Moreover, the adaptation of security mechanisms regarding Text and Data Mining (TDM) technologies is described as a technological option, focusing on the TDM deployed by the National Library of Greece alongside some considerations for applied Internet Security solutions that take into account GDPR requirements.

Furthermore, the implementation of IT infrastructures and websites with regards to cybersecurity and the GDPR is promoted. This sets the standards for compliant infrastructures and entities with regulations but also maintains them strong and resilient against most cyber threats.

Lastly, in the context of examining security issues regarding Internet access from public Wi-Fi networks, a profound evaluation and analysis of a corpus of approximately one million collected passwords from Wi-Fi networks was conducted for the first time. The data collected are compared against private password databases from previous research, in order to identify similarities and differences, underlying how convenience outweighs consequence, with a special reference to how people use their mobile devices, and consequently, explaining security naivety.

Περίληψη

Εκατομμύρια εταιρείες και δισεκατομμύρια χρήστες παγκοσμίως βασίζονται σε ασύρματα ή ενσύρματα δίκτυα για την καθημερινή τους εργασία και διασκέδαση. Λόγω της απουσίας ισχυρών μηχανισμών ασφαλείας κατά τον σχεδιασμό, υπάρχουν πολλοί τρόποι διείσδυσης κακόβουλων χρηστών σε δίκτυα και συστήματα. Το πανταχού παρόν πλέον παγκόσμιο διαδίκτυο, το οποίο έχει γίνει πιο φορητό και προσβάσιμο από ποτέ, μέσω ιδιωτικών και δημοσίων διαθέσιμων υποδομών πληροφορικής, καθιστά την μη εξουσιοδοτημένη πρόσβαση πιο εφικτή. Αυτό δημιουργεί επίσης σοβαρές ανησυχίες για την ασφάλεια και το απόρρητο λόγω πολλαπλών απειλών στον κυβερνοχώρο που προκύπτουν, ειδικά σε περίπτωση πρόσβασης στο Διαδίκτυο μέσω δημόσιων δικτύων Wi-Fi. Στο πλαίσιο αυτό, η διαδικτυακή ασφάλεια πρέπει και μπορεί να διαδραματίσει σημαντικό ρόλο στην προστασία της καθημερινής ζωής και των διαδικτυακών αλληλεπιδράσεων. Ωστόσο, οι περισσότεροι χρήστες δεν γνωρίζουν αυτές τις απειλές και το βαθμό στον οποίο το απόρρητό των επικοινωνιών ενδέχεται να τεθεί σε κίνδυνο.

Κανονισμοί, όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (ΓΚΠΔ), έχουν θεσπιστεί για τη διαφύλαξη και τη βελτίωση της ιδιωτικής ζωής και της ασφάλειας των χρηστών και των υποδομών πληροφορικής, επιβάλλοντας την εγκατάσταση κατάλληλων μέτρων ασφάλειας στον κυβερνοχώρο. Τέτοιοι κανονισμοί μπορούν να λειτουργήσουν ως ένα χρήσιμο σύνολο εργαλείων, το οποίο, μεταξύ άλλων απαιτήσεων, επιβάλλουν την υιοθέτηση του απορρήτου (και της ασφάλειας) κατά τον σχεδιασμό. Παρόλο που ο ΓΚΠΔ συνεπάγεται ένα ελάχιστο σύνολο τεχνικών μέτρων ασφαλείας που πρέπει να λαμβάνονται υπόψη από εταιρείες και οργανισμούς για την επίτευξη συμμόρφωσης, επισημαίνεται η ανάγκη υιοθέτησης ισχυρών μηχανισμών ασφαλείας που όχι μόνο θα θέσουν τις εταιρείες σε συμμόρφωση με κανονισμούς όπως ο ΓΚΠΔ αλλά θα διατηρήσουν παράλληλα ισχυρές και ανθεκτικές υποδομές έναντι πολλαπλών απειλών στον κυβερνοχώρο.

Στην παρούσα διατριβή, αναλύεται ένα μεγάλο σύνολο θεμάτων σχετικά με την προστασία της ιδιωτικής ζωής και της ασφάλεια, προσφέροντας λύσεις στα πολυάριθμα προβλήματα που αντιμετωπίζουν οι εταιρείες και οι μεμονωμένοι χρήστες είτε στην εργασία είτε σε ένα ψυχαγωγικό περιβάλλον καθημερινά. Ως μελέτη περίπτωσης, εξετάστηκαν σε

βάθος θέματα ασφάλειας και απορρήτου πληροφορικής σχετικά με την Εθνική Βιβλιοθήκη της Ελλάδας και το Δίκτυο Ελληνικών Βιβλιοθηκών της Εθνικής Βιβλιοθήκης της Ελλάδας, με βάση διεθνή κανονιστικά πρότυπα και πρότυπα ασφάλειας πληροφορικής. Επιπλέον, η προσαρμογή των μηχανισμών ασφαλείας σχετικά με τις τεχνολογίες Εξόρυξης Κειμένων και Δεδομένων (Text and Data Mining) περιγράφεται ως μια τεχνολογική επιλογή, εστιάζοντας στο TDM που αναπτύχθηκε από την Εθνική Βιβλιοθήκη της Ελλάδας παράλληλα με ορισμένες εκτιμήσεις για εφαρμοσμένες λύσεις ασφάλειας που λαμβάνουν υπόψη τις απαιτήσεις του ΓΚΠΔ.

Επιπλέον, προωθείται η υλοποίηση υποδομών πληροφορικής και ιστότοπων σχετικά με την ασφάλεια στον κυβερνοχώρο και τον ΓΚΠΔ, μέσω του οποίου τίθενται τα πρότυπα για τη συμμόρφωση σε επίπεδο ασφάλειας ενώ παράλληλα προτείνεται η διατήρηση ισχυρών και ανθεκτικών υποδομών πληροφορικής έναντι των περισσότερων απειλών στον κυβερνοχώρο.

Τέλος, στο πλαίσιο της εξέτασης ζητημάτων ασφάλειας σχετικά με την πρόσβαση στο Διαδίκτυο από δημόσια δίκτυα Wi-Fi, πραγματοποιήθηκε για πρώτη φορά στη βιβλιογραφία αξιολόγηση και ανάλυση ενός συνόλου περίπου ενός εκατομμυρίου συλλεγμένων κωδικών πρόσβασης από δίκτυα Wi-Fi. Τα δεδομένα που συλλέχθηκαν συγκρίθηκαν με στοιχεία πρόσβασης από ιδιωτικές βάσεις δεδομένων από προηγούμενες έρευνες, προκειμένου να εντοπιστούν ομοιότητες και διαφορές, καθώς και περαιτέρω χρήσιμα συμπεράσματα αναφορικά με τον τρόπο με τον οποίο οι άνθρωποι χρησιμοποιούν τις κινητές συσκευές τους που είναι προσβάσιμες στο διαδίκτυο.

INDEX

LIST OF TABLES

Chapter 4

Chapter 5

LIST OF FIGURES

Chapter 3

Chapter 1: Introduction

As the world has become ever more mobile, the need to stay connected, either to maintain personal relationships or do business, is growing for a large number of people. While, in the past, this was done mainly on our personal computers, networking today with the internet of things, has become more portable and accessible than ever before through mobile and smart devices that we carry with us all day and use on public networks.

Global public Wi-Fi networks grew from 1.3 million in 2011 to 5.8 million in 2015, a 350% increase in just four years, responding to our need for instant information and in the convenience of our own apparatus (Balasubramanian et al., 2009). Public networks, offered in public places such as cafeterias and airports or by municipalities, allow the users to connect their devices using weak passwords or no passwords at all (Tanshkova, 2020). As of 2017, around 59% of people from a total corpus of 15,532 stated that they had logged in to a personal email account over public Wi-Fi (Clement, 2019), while 56% stated that they had logged in to their social media accounts. IT professionals tend to advise against using these public networks for tasks that require sensitive personal information as it may be accessible by other users of the network. Public opinion tends to consider public Wi-Fi networks as secure: 61% of people that responded to the survey, stated that they feel safe on a public Wi-Fi network, while 39% state that they feel unsafe (Clement, 2017).

According to a global study of 1,800 college students and young professionals, 29% check their phones so constantly that they lose count, and one in five at least every 10 minutes. Indeed, we rely on our smart devices instinctively, hungry. In the age of ubiquitous computing, we often seek convenience and speed.

The craving to stay connected indicates that this will be done from anywhere, blending and overlapping public and private spaces. Yet, even though public networks are a welcome service to many people, they pose significant concerns regarding the privacy and the security of the data transmitted within such networks. What happens when users enter the public domain? In other words, what is out there for users? While more and more governments promise to widen the availability of public networks, and as public Wi-Fi hotspots are everywhere in big cities, awareness of security issues does not seem to follow a parallel path.

As staying connected has become a daily routine for most people, the number of ensuing security threats is not a matter of concern neither for the users themselves nor for the administrators. Nowadays, a malicious user does not need to be adept at social engineering to get unauthorized access to a network or obtain by illegitimate means a Wi-Fi password (Krombholz et al., 2015). Regardless of how easy it has become for an attacker to achieve unauthorized access to a network or service (Chou et al., 2013; Kelley et al., 2012), security and privacy are more than often compromised over the exigency to implement IT infrastructures without proper security mechanisms in order to facilitate the usability and speed of the service. This lack of security affects the protection and the privacy of the data within all kinds of infrastructures such as websites, wired and wireless networks.

While the literature abounds with references to attacks occurring in private Wi-Fi infrastructures (WarDriving, Dictionary attacks, Password Theft, and WEP/WPA or WPS attacks), the dangers in public networks are somewhat different as the password is already known. Indeed, uncontrolled access to public Wi-Fi hotspots and robust mobile security often conflict with one another. Users not only run the risk of malware, Trojan, or ransomware infections, but any password or login credentials they enter are transmitted in cleartext due to the lack of encryption, making them ideal targets for cybercriminals. Businesses also face numerous problems such as the proliferation of legitimate-looking twin networks set up by cybercriminals to steal information and jeopardize further the privacy and security of users and data within the infrastructure. Apart from the actual breach, companies face enormous fines due to regulations on a global scale, such as the GDPR, in cases of data leakages and cyber-attacks. The GDPR offers a digital environment for companies and organizations where they can better trace, secure, and handle data within the IT infrastructure and beyond.

1.1 Research Contribution and Structure

This thesis describes the adaptation of security mechanisms regarding Text and Data Mining technologies, the implementation of IT infrastructures and websites with regard to cybersecurity and the GDPR, and the challenges of open Wi-Fi networks.

The second chapter explores Text and Data Mining (TDM) as a technological option, focusing on the TDM deployed by the National Library of Greece (NLG) and considerations for applied Internet Security solutions taking into account the GDPR requirements. TDM is usually leveraged upon by large libraries worldwide in the technologically enhanced processes of web-harvesting and web-archiving with the aim to collect, download, archive, and preserve content and pieces of work that are found available on the Internet. TDM is used to index, analyze, evaluate, and interpret mass quantities of data including texts, sounds, images, or data through an automated "tracking and pulling" process of online material. Access to the web content and works available online are subject to restrictions by legislation, especially to laws about Copyright, Industrial Property Rights, and Data Privacy. As far as Data Privacy is concerned, the application of the GDPR is considered an issue of vital importance for the smooth operation of TDM services offered by national libraries mostly in the EU Member States, which among other requirements mandates the adoption of privacy-by-design and advanced security techniques.

Taking into account the requirements of the GDPR, the third chapter analyzes their implementation with regard to applied Internet Security solutions. While the Regulation offers a minimum set of technical Internet Security means to be taken into consideration by companies and organizations to achieve compliance with the GDPR, this chapter highlights the adaptation of strong security mechanisms that will not only set compliant infrastructures and entities with the GDPR but also maintain them strong and secure against most threats.

In a similar vein with the previous part, chapter four will cover the implementation of requirements of the GDPR concerning applied Internet Security solutions for the websites. The GDPR offers a minimum set of technical Internet Security means to be taken into consideration by companies and organizations Europe-wide to achieve the GDPR compliance. This chapter analyzes the adaptation of strong and proper security mechanisms that will not only set compliance with the GDPR for library-members of the Greek Libraries Network of the NLG but will also maintain them strong and secure against most cybersecurity threats, both internal and external, targeting websites.

Understanding the present moment as part of a longer lifecycle regarding passwords and their use in the daily life, the fifth chapter will present the collection for the first time of

a uniquely compiled vast database of approximately one million passwords from Wi-Fi networks from the biggest cities and capitals around the world, employing an innovative methodology. In this chapter, the passwords of publicly available networks are examined, by analyzing their complexity and strength. Public Wi-Fi networks, while certainly useful and increasingly available, have numerous pitfalls. On their own, they pose significant security and privacy concerns for users, but in combination with a lack of understanding about the risks, the threats are markedly amplified. Second, the data collected are compared against private password databases from previous research in order to identify similarities and differences. Based on this analysis, the solutions offered in the literature are evaluated and a number of recommendations are proposed to raise security awareness. In doing so, new insights are provided on how businesses who offer public Wi-Fi networks perceive security and ultimately deal with it. Furthermore, by comparing this database with other private password datasets, this research aims for the first time in the literature to lay down a more rounded idea of how convenience outweighs consequence, especially with reference to how people use their mobile devices; thus, explaining security naivety.

This is both necessary and timely. First, it is necessary because the literature has tended to focus mainly on analyzing private password databases. Second, it is timely because, in the age of ubiquitous computing when people seek to stay connected at all costs, there is a high need for discussion of alternatives for the enhancement of the security of public Wi-Fi networks. This is an important time to conduct new research that focuses on massively used networks across the world.

Chapter 2: Text and Data Mining for the National Library of Greece in consideration of Internet Security and the GDPR

2.1 Overview

Text and Data Mining (TDM) as a technological option is usually leveraged upon by large libraries worldwide in the technologically enhanced processes of web-harvesting and web-archiving with the aim to collect, download, archive, and preserve content and works that are found available on the Internet. TDM is used to index, analyze, evaluate, and interpret mass quantities of works including texts, sounds, images, or data through an automated 'tracking and pulling' process of online material. Access to the web content and works available online are subject to restrictions by legislation, especially to laws pertaining to Copyright, Industrial Property Rights, and Data Privacy. As far as Data Privacy is concerned, the application of the General Data Protection Regulation (GDPR) is considered as an issue of vital importance for the smooth operation of TDM service offered by national libraries mostly in the EU Member States. The GDPR, among other requirements, mandates the adoption of privacy-by-design and advanced security techniques. This article focuses on the TDM deployed by the National Library of Greece (NLG) and the considerations for applied Internet Security solutions taking into account the GDPR requirements. NLG has deployed TDM as of February 2017 in consideration of the provision of Art.4(4)(b) of law 4452/2017, as well as of the provisions of Regulation 2016/679/EU (GDPR). Art.4(4)(b) of the Law 4452/2017 sets the TDM activity in Greece under the responsibility of NLG, appointed as the organization to undertake, allocate and coordinate the action of archiving the Hellenic web, i.e. as the organization responsible for the text and data analysis at a national level in Greece. The deployment of TDM by NLG, presented by the authors, caters for a framework of technical and legal considerations so that the electronic service enabled based on the TDM operation complies with the data protection requirements set by the new EU legislation. While the presentation elaborates upon the minimum set of technical Internet Security means considered by NLG for achieving GDPR compliance, this chapter focuses on TDM and GDPR issues specifically in relation to Art.89 of the GDPR titled '*Safeguards and derogations relating to processing for archiving purposes in the public*

interest, scientific or historical research purposes or statistical purposes' that is a key term ruling for the operation of NLG in compliance with the GDPR.

2.2 Text and Data Mining under GDPR

2.2.1 Text & Data Mining and GDPR issues for the National Library of Greece

Text and Data Mining (hereinafter, TDM) activity may involve the processing of personal data. This processing of personal data, though, is processing for archiving purposes in the public interest, or processing for scientific or historical research or statistical purposes. In many cases it is processing that combines more than one of the above-mentioned purposes.

Under the General Data Protection Regulation (Regulation 2016/679/EU, hereinafter GDPR)[1], the data protection principles set out the main responsibilities for organizations. These principles are applicable in the case of organizations that benefit from the TDM exception, of course. The principles are similar to those described in Directive 95/46/EC (the Data Protection Directive)[2] of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. However, the GDPR has added details at certain points and a new accountability requirement. The most significant addition is the accountability principle: the GDPR demands from a data processor to show how it complies with the data protection principles, for example by documenting the decisions it takes about a processing activity.

The data protection principles are described in Article 5 of the GDPR[3, 4], which lays down all the key principles for the protection of personal and special categories data, i.e. the

[1] Regulation 2016/679/EU of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), available at URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [last check, April 30, 2020].

[2] The Data Protection Directive (officially Directive 95/46/EC) on the protection of individuals with regard to the processing of personal data and on the free movement of such data is a European Union directive adopted in 1995, which regulated the processing of personal data within the European Union (EU). The General Data Protection Regulation has superseded the Data Protection Directive and came into force as of May 25, 2018.

[3] See, also, Data Protection Directive, Art.4 titled '*Principles relating to processing of personal data*'. See, also, Recital 39 according to which "*Any processing of personal data should be lawful and fair. It should be transparent to natural*

lawfulness, fairness, transparency, purpose-limitation, data-minimization, accuracy, storage limitation, integrity and confidentiality, and accountability. According to the provision of Article 5 of the GDPR:

1. *Personal data shall be:*

   a. processed lawfully[5]*, fairly*[6]*, and in a transparent*[7] *manner in relation to individuals*[8]*;*

---

persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Every reasonable step should be taken to ensure that personal data which are inaccurate are rectified or deleted. Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including for preventing unauthorized access to or use of personal data and the equipment used for the processing."

[4] See, also, Directive 2016/680/EU Art.4 for principles relating to processing of personal data, and Recitals 26-28 in this Directive.

[5] The lawfulness of the processing is described in Art.6(1) of the GDPR; relevant to the lawfulness of the processing are Recitals 40-49. The conditions of data subject's consent are described in Art.7 of the GDPR. Regarding a child's consent in the case of offering of Information Society services, Art.8 of the GDPR applies. Lawfulness of the processing means that personal data processing respects all applicable requirements; personal data processing should be considered as lawful if processing is in accordance with law, pursues a legitimate purpose, and is necessary and proportionate in a democratic society in order to achieve that purpose.

[6] Fair processing implies that personal data or special categories data have not been obtained or otherwise processed through unfair means, by deception or without the knowledge of data subject.

[7] Transparency means that it should be clear to natural persons that personal data concerning them are collected, used, consulted or otherwise processed. Relevant is Recital 39, which explains transparency and sets requirements for the quality of information to be given to data subjects: it should be easily accessible and easy to understand, and it should also include information through which natural persons should be made aware of risks and safeguards in relation to the processing of their personal data. See, also, Art.12 of the GDPR; relevant to the transparency of the processing are Recitals 58-59.

[8] See, also, Art.8 of Directive 2016/680/EU.

b. collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes[9]*;*

c. adequate, relevant, and limited to what is necessary, in relation to the purposes for which they are processed[10]*;*

d. accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay[11]*;*

e. kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods in so far as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to the implementation of the appropriate technical and organizational measures required by the GDPR in order to safeguard the rights and freedoms of individuals[12]*;*

f. processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against

---

[9] The purpose-limitation principle requires data to be processed for specified, explicit, and legitimate purposes (the purpose-dimension of this principle) and not further processed in a manner that is incompatible with those purposes (the compatible-dimension of this principle). Both dimensions of the purpose-limitation principle should occur at the time of collection of the personal data, i.e. at the beginning of the processing of personal data and/or special categories data. Relevant to the purpose-limitation principle is the provision of Art.6(4) of the GDPR. There are only two cases for exception of the purpose-limitation principle: 1) if the data subject consents to a new, incompatible purpose for his/her data processing, and 2) if the processing is based on EU or Member-State law. Aside from these two cases, the GDPR considers as a priori compatible with the initial purpose of data processing the cases of processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

[10] The data-minimization principle pertains to both the quantity and quality of data, which should only be processed only if the purposes aimed cannot be fulfilled by other means. Recital 39 is relevant to the minimization-principle.

[11] The accuracy principle. See, also, Art.7(2) of Directive 2016/680/EU.

[12] The storage-limitation principle. See, also, Art.25 of the GDPR and Art.20 of Directive 2016/680/EU. Relevant are Art.4(1)(e) and Art.5 of Directive 2016/680/EU.

accidental loss, destruction, or damage, using appropriate technical or organizational measures[13].

2. *The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability')*[14].

The processing of personal data through TDM activity is inevitable. TDM involves the processing of text and data which may include any information relating to an identified or identifiable natural person, a.k.a. personal data[15]. Essential to the concept of personal data is the linkability of information to an individual allowing his/her identification. Regarding TDM and personal data protection, there is concern that sets of correlated data, which could be considered insignificant or even trivial, can provide intimate knowledge about data subjects where TDM is applied (Hargreaves et al., 2014). Any information that allows for the identification of a natural person by reasonable means may constitute personal data. Truly anonymous data do not constitute personal data, as is stated in Recital 26 of the GDPR: "[...]*The principles of data protection should therefore not apply to anonymous information, namely, information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*."

TDM constitutes processing of personal data, in the sense that it involves any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction[16].

In most cases TDM works in the following manner (Botti et al., 2019b):

1. It identifies input materials to be analyzed, such as works, or data individually collected or organized in a pre-existing database;

[13] The integrity and confidentiality principle. Art.32-34 of the GDPR are relevant to this principle. Also, Art.4(1)(f) and Art.39-31 of Directive 2016/680/EU focus on the integrity and confidentiality principle.

[14] The accountability principle means that the controller must be able to demonstrate that the processing is in compliance with the legal applicable rules. Relevant is Art.24 of the GDPR.

[15] Art.4(1) of the GDPR.

[16] Art.4(2) of the GDPR.

2. It copies substantial quantities of materials—which encompasses:

a. pre-processing materials by turning them into a machine-readable format compatible with the technology to be deployed for the TDM so that structured data can be extracted. Pre-processing typically encompasses the following tasks:

i. *Tokenization*: This is typically the first step in a natural language processing solution, and it refers to splitting the text into meaningful character sequences/self-contained semantic units, e.g. words or sentences.

ii. *Normalization*: This involves removing morphological variations from words such as capitalization, plural number or tenses, in order to grasp similarities between them (e.g., the same word in singular and plural), obviously with a loss of information. Two types of techniques are used regarding normalization. These are *stemming* and *lemmatization*. In the former, language-specific patterns are recognized, using for example the rules for converting words from singular to plural or verb tenses. This technique is simple, fast, and applicable for large volumes of text. Lemmatization involves using a dictionary (such as WordNet that is both a dictionary and a thesaurus) to extract the roots of common words. This approach can be more accurate compared to stemming, but it is more resource-intensive, and dictionaries may be incomplete for certain languages. The two methods can complement each other, and they are often used in conjunction.

iii. *Parsing*: This involves a group of functions that are used after term isolation and document cleanup, i.e., after normalization and parsing, which facilitate working in higher abstraction layers. Typically, parsing includes morphological and syntactical analysis of tokens in order to identify their role within sentences (e.g. noun, verb, adjective, or object-verb-subject), which is referred to as *Part-of-Speech (POS) tagging*.

b. possibly, but not necessarily, uploading the pre-processed materials on a platform, depending on the TDM technique to be deployed;

3. It extracts the data; and

4. It recombines data to identify patterns into the final output.

Therefore, to undertake TDM a researcher must access and make a copy of the work/data in order to apply the necessary algorithms for the extraction of new knowledge. This necessary

copying of the work/data in the process of the application of TDM has led to considerations of the necessity for an open norm in the European Copyright legal framework which could be similar to the open norm of the '*fair use*' doctrine in the American Copyright Act. Unfortunately, there's no room for such an open norm doctrine in the EU Copyright law, for the time being (Botti et al., 2019a; Botti et al., 2019b).

It has also led to considerations regarding data protection which have become more vivid taking into account the '*straitjacket*' of the GDPR. TDM is restricted by the GDPR. It is not an activity that can be lawfully executed without restrictions. The application of the GDPR rules on TDM restricts the later through the principles of processing, the legal grounds for the processing, and informatory obligations of the data subjects for the processing, at least (Caspers et al., 2016). Thus, the collection and processing of personal data in the framework of TDM activity undertaken for scientific research purposes are subject to the safeguards imposed by the GDPR principles, such as the necessity of having a legitimate ground to process such data, the obligation to collect data only as far as it is necessary in order to achieve the specified and legitimate purpose (principle of finality/the purpose limitation principle of Art.5(1)(b) of the GDPR)[17]; the prohibition against collecting more data and to keep them for a longer period than is necessary for the purposes for which they are collected and/or further processed (the 'data minimization' principle)[18]. Also, the organization which deploys TDM is bound by the principle of accountability which means that said organization must be able to demonstrate that it has appropriate processes to ensure that it only collects and holds the personal data that it needs in order to achieve the scientific purpose for which TDM was deployed. Besides, said organization must bear in mind that the GDPR says individuals have the right to complete any incomplete

[17] According to Art.5(1)(b) of the GDPR personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes. In practice, the purpose limitation principle means that the organization which deployed TDM must: (1) be clear from the outset why it is collecting personal data that may be included in the text or data aimed to be mined and what it intends to do with it; (2) comply with said organization's documentation obligations to specify the purposes for TDM which may involve the collection of personal data; (3) comply with the organization's transparency obligations to inform individuals about its purposes regarding TDM and the processing of personal data mined; and (4) ensure that if the organization plans to use or disclose personal data for any purpose that is additional to or different from the originally specified purpose, the new use is fair, lawful and transparent.

[18] According to Art.5(1)(c) of the GDPR personal data shall adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimization). This means that the organization which deploys TDM activity must identify the minimum amount of personal data that it needs to fulfil the purpose of scientific research through TDM. Once this minimum amount of personal data is identified, the organization should hold no more personal data than what was identified as necessary.

data which is inadequate for the organization's purpose, under the right to rectification[19]. They also have the right to force the organization which processed their data to delete any data that is not necessary for the organization's purpose, under the right to erasure (the right to be forgotten)[20].

2.2.2 Article 89 of the GDPR

In the case of TDM activity undertaken by the National Library of Greece (hereinafter, NLG) Article 89 of the GDPR is applicable. According to Article 89 of the GDPR, titled '*Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*' (emphasis through underscore added by the authors):

1. Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation (Regulation 2016/679/EU), for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organizational measures are in place, in particular in order to ensure respect for the principle of data minimization. Those measures may include pseudonymization provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

2. Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15[21], 16[22], 18[23], and 21[24] subject to the conditions and safeguards

[19] Art.16 of the GDPR.

[20] Art.17 of the GDPR.

[21] Right of access by the data subject

[22] Right to rectification

[23] Right to restriction of processing

referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

3. Where personal data are processed for archiving purposes in the public interest, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18, 19[25], 20[26], and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

4. Where processing referred to in paragraphs 2 and 3 serves at the same time another purpose, the derogations shall apply only to processing for the purposes referred to in those paragraphs.

Article 89(1) of the GDPR repeats the principles[27] of data minimization[28], purpose limitation[29], and storage limitation[30]. The whole Article 89 encompasses the processing of both

---

[24] Right to object

[25] Notification obligation regarding rectification or erasure of personal data or restriction of processing

[26] Right to data portability

[27] See Article 4 of Directive 2016/680/EU regarding the principles relating to the processing of personal data. See, also, Recital 26 of the same Directive, according to which "*Any processing of personal data must be lawful, fair and transparent in relation to the natural persons concerned, and only processed for specific purposes laid down by law. … Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of their personal data and how to exercise their rights in relation to the processing. In particular, the specific purposes for which the personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate and relevant for the purposes for which they are processed. It should, in particular, be ensured that the personal data collected are not excessive and not kept longer than is necessary for the purpose for which they are processed. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Member States should lay down appropriate safeguards for personal data stored for longer periods for archiving in the public interest, scientific, statistical or historical use."*

[28] Art.5(1)(c) of the GDPR. The principle of data minimization is a specification of the general principle of proportionality. The principle of data minimization posits that the collection of personal data shall be "[…] *adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization')."*

[29] Art.5(1)(b) of the GDPR. According to this principle, personal data may be "[…] *collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical*

personal data of Article 6 of the GDPR as well as special categories data of Article 9 of the GDPR[31]. Article 89 does not describe a legal basis for the processing of personal data or of special categories data. The legal bases for the processing of this data are described strictly in Article 6 of the GDPR. Article 6(1) of the GDPR exhaustively stipulates what may constitute a legal basis for data processing. Therefore, the processing of data through the TDM application in a library setting could be lawful only if either one of the options described in Article 6(1) of the GDPR is applicable. Most likely, processing of data in the framework of TDM complies with processing that is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child[32].

Paragraphs 2 and 3 of Article 89 entitle the Member States to provide for derogations from certain rights of the data subjects. The scope of Article 89(2) is limited to processing for scientific or historical research purposes and statistical purposes. Said paragraph of Article 89 provides for derogations from the right of access[33], the right of rectification[34], the right of restriction of processing[35], and the right of objection[36]. However, such derogations are still subject to the conditions and safeguards referred to in Article 89(1) of the GDPR, which means

---

purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')."

[30] Art.5(1)(e) of the GDPR. According to this principle, personal data may be "[…] *kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods in so far as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation')."* Upon expiration of that period, data must be deleted or anonymized.

[31] See Art.9(2)(j) of the GDPR according to which "*processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject."*

[32] Art.6(1)(f) of the GDPR.

[33] Art.15 of the GDPR.

[34] Art.16 of the GDPR.

[35] Art.18 of the GDPR.

[36] Art.21 of the GDPR.

firstly that appropriate safeguards must be in place to protect the rights and freedoms of data subjects even when derogations apply[37, 38, 39]; secondly, that the use of the rights from which derogations are given must be likely to render impossible or seriously impair the achievements of the specific purposes, and such derogations are necessary for the fulfillment of those purposes[40]; and thirdly, that derogations only apply to the purposes mentioned in the respective paragraphs of Article 89.

Recital 159 states that "*the processing of personal data for scientific research purposes should be interpreted in a broad manner including, for example, technological development and demonstration, fundamental research, applied research and privately funded research.*"

---

[37] See Recital 156 according to which "[…] *the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation. Those safeguards should ensure that technical and organizational measures are in place in order to ensure, in particular, the principle of data minimization. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data). Member States should provide for appropriate safeguards for the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Member States should be authorized to provide, under specific conditions and subject to appropriate safeguards for data subjects, specifications and derogations with regard to the information requirements and rights to rectification, to erasure, to be forgotten, to restriction of processing, to data portability, and to object when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. The conditions and safeguards in question may entail specific procedures for data subjects to exercise those rights if this is appropriate in the light of the purposes sought by the specific processing along with technical and organizational measures aimed at minimizing the processing of personal data in pursuance of the proportionality and necessity principles."*

[38] See Recital 157 according to which "*In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law."*

[39] See Recital 162 according to which "*Union or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality."*

[40] See Art.14(5)(b) of the GDPR, according to which the provision of information where personal data have not been obtained from the data subject shall not apply in so far as "[…] *the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available."* See, also, Art.17(3)(d) of the GDPR, according to which the provisions of paragraphs 1 and 2 of Article 17 that pertains to the right to erasure (the right to be forgotten) shall not apply to the extent that processing is necessary "[…] *for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing."*

National legislation of the Member-States may include a definition of the term '*scientific research purposes*.' The intention of the EU legislator is to include under the definition of '*scientific research purposes*' the broadest possible meaning and allow scientific research purposes to be pursued, at least to the extent possible under the Data Protection Directive. In the Greek legal framework, there is no specific definition of the term '*scientific research purposes*', although the law 3653/2008 pertains to the enhancement of scientific research and technology in Greece. The individual right to scientific research and teaching is recognized in the wording of Article 16(1)(a) of the Greek Constitution according to which '*Art and science, research and teaching shall be free and their development and promotion shall be an obligation of the State.*' However, so far there are very few Greek Court decisions elaborating upon this individual right and the meaning of '*scientific research purposes*', with the most notable being the decision of the Council of State No.1043/1989 (Papadopoulos, 2020).

The term '*historical research*' is not defined in the GDPR. However, Recital 160 makes clear that under the term '*historical research*' fit both '*historical research' and 'research for genealogical purposes*'.

Recital 162 defines the meaning '*statistical purposes*' as "*any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results.*" Further, the statistical purpose in the processing of data implies that the result of the processing is not personal data, but aggregate data that may not be used in support of measures or decisions regarding any particular natural person (data subject). Special regulation may be applicable in case of processing of data for statistical purposes, such as Article 338(2) of the Treaty for the Functioning of the European Union (TFEU)[41] or the Regulation EC 223/2009 on European Statistics[42].

[41] According to Art.338(2) of the TFEU "*The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators."*

[42] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland), OJ L 87, 31/03/2009, p.164–173.

Article 89 does not distinguish between research pursuing public interests and research for private and/or purely commercial purposes. Thus, it applies in research pursued through TDM or other means either for public or for private interest, either for commercial or for non-commercial purposes.

Article 89(3) applies to archiving purposes in the public interest. Not every archive falls under the scope of Article 89(3), but only those that have a legal obligation to maintain records in the scope of the public interest. According to Recital 158, "*Public authorities or public or private bodies that hold records of public interest should be services which, pursuant to Union or Member State law, have a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for the general public interest.*" This means that archives which do not fit in the public interest scope, are not covered by Article 89. NLG is entitled to hold records of public interest and has a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for general public interest according to Greek law 3149/2003 as amended through law 4452/2017.

Under Article 89(3) of the GDPR, where personal data are processed for archiving purposes in the public interest, EU law of Member-State law may provide for derogations from the right of access by the data subject[43], the right of rectification[44], the right to restriction of processing[45], the notification obligation[46], the right to data portability[47], and the right to object[48]. However, such derogations in the case of Article 89(3) are still subject to the conditions of Article 89(1) of the GDPR.

In the case of processing of special categories of personal data, i.e. personal data of Article 9 of the GDPR, either for archiving purposes in the public interest or for scientific and historical research purposes or for statistical purposes, national law may stipulate conditions for the

[43] Art.15 of the GDPR.

[44] Art.16 of the GDPR.

[45] Art.16 of the GDPR.

[46] Art.19 of the GDPR.

[47] Art.20 of the GDPR.

[48] Art.20 of the GDPR.

lawfulness of the processing, according to Article 9(2)(j) of the GDPR. Article 22 of the Greek law 4624/2019 caters for the lawfulness of the processing of special categories of personal data, Said Article of law 4624/2019 is applicable in the case of processing of special categories of personal data through the TDM process that is considered lawful according to Article 9(2)(a) of law 4624/2019, i.e. on the basis of (absolutely) necessary processing for the purpose of public interest.

In addition to the public interest purpose, TDM activities by the NLG may also be undertaken for historical[49] or scientific[50] research purposes, at least. According to Recital 159 of the GDPR, in order to meet the specificities of processing personal data for scientific research purposes, specific conditions must apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes. If the result of scientific research gives reason for further measures in the interest of the data subject, the general rules of Regulation 2016/679/EU should apply in view of those measures. Thus, given that the output of TDM deployed by the NLG includes personal data found in the works harvested from the Web, NLG must apply specific conditions regarding the publication or otherwise disclosure (in copyright terms, this is deemed to be relevant to the presentation/communication to the public right of copyright[51]) of the output of the TDM that includes personal data of persons who have not deceased. These specific conditions could pertain to the intranet or the extranet through which the TDM output is accessible to a certain public that is narrower than the general public.

Regarding TDM activities undertaken for statistical purposes, NLG must cater for the following requirements in consideration of Recital 162 of the GDPR: the result of the processing

[49] See Recital 160 of the GDPR according to which "*Where personal data are processed for historical research purposes, this Regulation should also apply to that processing. This should also include historical research and research for genealogical purposes, bearing in mind that this Regulation should not apply to deceased persons."*

[50] The 'scientific research purpose' is meant widely for the application of the GDPR. According to Recital 159, "*For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area. Scientific research purposes should also include studies conducted in the public interest in the area of public health."*

[51] See Art.3(1)(h) of the Greek Copyright law 2121/1993, which pertains to the communication to the public right of the works of copyright-holders, by wire or wireless means or by any other means, including the making available to the public of their works in such a way that members of the public may access these works from a place and at a time individually chosen by them.

of personal data or special categories data whenever is undertaken either in the framework of TDM or other means for statistical purposes cannot be personal data, but only aggregate data; the result of the processing for statistical purposes cannot be used in support of measures or decisions regarding any particular data subject (natural person). In consideration of Recital 162, statistical results may further be used for purposes other than scientific research purposes. The term '*statistical purposes*' is meant as "*any operation of collection and the processing of personal data necessary for statistical surveys or the production of statistical results*[52]." According to Recital 162 of the GDPR "*The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.*" Thus, there are two conditions, which both must be met in the case of processing of personal data or special categories data for statistical purposes: a) the result of the processing of personal data or special categories data must not be personal data but should be aggregate data; b) the result of the processing of personal data or special categories data must not be used in support of measures or decisions regarding any particular natural person.

According to Art.5(1)(b) of the GDPR, personal data processed during TDM activities undertaken by the NLG for historical or scientific purposes may further be processed for archiving purposes in the public interest or for statistical purposes, in accordance with Art.89(1), without being considered to be incompatible with the initial purposes ('*purpose limitation*'). Regarding the '*storage limitation*' requirement, Art.5(1)(e) of the GDPR rules that personal data processed during TDM activity may be stored for longer periods in so far as the personal data are processed solely for archiving purposes[53] in the public interest, scientific or historical research purposes, or statistical purposes in accordance with Art.89(1), and are subject to the implementation of appropriate technical and organizational measures required by Regulation 2016/679/EU in order to safeguard the rights and freedoms of the data subject.

[52] Recital 162 of the GDPR.

[53] According to Recital 158, the 'archiving purpose' includes in particular "*providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust or war crimes.*"

Regarding the processing of personal data for archiving purposes through the TDM process, data referring to deceased people do not constitute personal data[54], thus there is no conflict in processing them with the provisions of the GDPR. In the same vein, further processing of personal data for archiving purposes, for example with a view to providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust, or war crimes, is not in conflict with the provisions of the GDPR[55].

Art.89(1) makes a reference to '*pseudonymization*' as a technical measure to ensure respect for the principle of data minimization. Pseudonymization is meant as "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*[56]." Art.89(1) refers to '*pseudonymization*', but not to '*anonymization*'. However, the reference in Art.89(1) to "*further processing which does not permit or no longer permits the identification of data subjects*" could be interpreted as including anonymization. This interpretation is also inferred from Recital 156, according to which "*The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfill those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data).*" The list of safeguards mentioned in Art.89(1) is not exhaustive, thus both anonymization and pseudonymization could be favored under Article 89 of the GDPR. The legal distinction between anonymized and pseudonymized data is the categorization as personal data. Pseudonymous data still allows for some form of re-identification (even indirect and remote),

[54] See Art.4 and Recital 158 of the GDPR.

[55] See Recital 158 of the GDPR, according to which "*Member States should also be authorized to provide for the further processing of personal data for archiving purposes, for example with a view to providing specific information related to the political behavior under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust, or war crimes.*"

[56] Art.4, No.5 of the GDPR.

while anonymous data cannot be re-identified[57]. Anonymized data are not subject to data protection obligations through the application of the GDPR or relevant legislation. Anonymous data cannot be linked back to identifiable data subjects. Also, anonymous data is useless for almost anything but high-level data aggregation and analysis[58].

In contrast to anonymized data, pseudonymized data retains some statistical utility relative to the level of pseudonymization. For this reason, data pseudonymization is preferable for statistical analysis. That is why Art.89(1) of the GDPR refers to '*pseudonymization*' rather than to '*anonymization*' as a technical measure to ensure data minimization. Pseudonymization techniques differ from anonymization techniques. With anonymization, the data is scrubbed for any information that may serve as an identifier of a data subject. Pseudonymization does not remove all identifying information from the data but merely reduces the linkability of a dataset with the original identity of an individual (e.g., via an encryption scheme). Both pseudonymization and anonymization are encouraged in the GDPR and enable its constraints to be met. These techniques should therefore be generalized and recurring. Those in possession of personal data should implement one of these techniques to minimize risk, and automation can reduce the cost of compliance.

Pseudonymization in the GDPR is referred to as a method—a technical means—that can be used for demonstrating GDPR compliance in more than one Articles or Recitals of the Regulation. Article 25(1) of the GDPR makes a reference to pseudonymization as an appropriate technical measure which is designed to implement data-protection principles. Article 32(1)(a) of the GDPR names pseudonymization and encryption of data as a technical means to ensure a level of security appropriate to the risk, thus pseudonymization is advocated as a risk-based approach to data security. Also, Recital 78 of the GDPR reports pseudonymization of personal

[57] See Recital 26 of the GDPR according to which *"[…] Personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. […] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*"

[58] See Working Party of Article 29, *Opinion 05/2014 on Anonymization Techniques*, WP 216, April 10, 2014, available at URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [last check, April 30, 2020].

data as soon as possible as a measure that meets the principles of data protection by design and by default.

There are multiple methods for pseudonymization such as data masking, encryption, and tokenization. Encryption entails the use of a key to encode or protect a data set. Consequently, encryption is mathematically reversible and is subject to the complexities of key management. Tokenization, by comparison, involves replacing identifying or sensitive data with a mathematically unrelated value. Therefore, the tokens cannot be mathematically reversed. Both encryption and tokenization can be format-preserving and tokens may optionally include elements of the original value for purposes of data processing. Tokenization is used in the TDM process. Data masking is a process for obfuscating data that is typically accomplished via encryption. Using masking, data can be de-identified and de-sensitized so that personal information remains anonymous in the context of support, analytics, testing, or outsourcing.

The most suitable method of pseudonymization depends on the specific use case and needs of an organization, although it is worth noting that from a compliance standpoint, tokenization via a cloud-based tokenization provider is the only method that enables an organization to completely remove sensitive or identifying data from its systems. This is a significant differentiator from both a compliance and a data security perspective. As is already stated above, tokenization is part of the TDM process.

TDM activities involve both web harvesting and web archiving processing of subject matter. Regarding the NLG, which is a public law entity according to Art.1(1) of law 3149/2003, thus a legal entity that aims at serving the general public interest, Recital 158(2) of the GDPR is of interest. NLG as a public body that holds records of public interest—these include the output of TDM records as well as works in the general or specific collections of works archived and made available to the public through NLG—is an organization that provides services which, pursuant to the Greek law, has a legal obligation to acquire, preserve, appraise, arrange, describe, communicate, promote, disseminate and provide access to records of enduring value for the general public interest.

Also, in consideration of Recitals 156 and 158 of the GDPR, NLG which is empowered to deploy TDM for archiving and research purposes for the public interest must cater for the application of appropriate safeguards for the rights and freedoms of the data subjects pursuant to

the GDPR. Those safeguards should ensure that technical and organizational measures are in place in order to particularly ensure the principle of personal data minimization or the principle of storage limitation. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes is to be carried out when the Controller—it might be a different entity than the NLG which is the Processor definitely—has assessed the feasibility to fulfill those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymization of the data).

Article 89(2) of the GDPR allows for derogations from rights referred to in Article 15 (the right of access), Article 16 (right of rectification), Article 18 (right of restriction of processing), and Article 21 (right of objection) of the GDPR when personal data are processed for scientific or historical research or statistical purposes in so far, as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes. Said GDPR article also posits that EU or Member State law may provide for derogations from the rights of the aforesaid articles. Regarding the Greek law, Article 29 and Article 30 of the law 4624/2019 are applicable in the case of NLG.

Derogations from the right to object (Art.21 of the GDPR) where personal data are processed for scientific or historical purposes or statistical purposes pursuant to Article 89(1) of the GDPR are provisioned in Article 21(6) of the GDPR, too. Article 21(6) allows for derogations from the right to object in cases of processing of personal data which is necessary for the performance of a task carried out for reasons of public interest.

Regarding the obligation to inform the data subject when personal data processed through TDM process have not been obtained from the data subject[59], Article 23(1) of the GDPR allows EU or Member State law to provide restrictions to the obligation to inform. As far as the Greek law is concerned, Article 32(1)(a)(aa) of law 4624/2019 is applicable in the case of NLG and its obligation to inform the data subjects for the processing of their personal data through TDM process. Thus, NLG is not obliged to inform the data subjects for the processing of their

[59] See Art.14 of the GDPR that pertains to information to be provided to the data subject where personal data are processed without being obtained from the data subject.

personal data through the deployment of TDM by NLG that serves important objectives of general public interest.

Finally, regarding the right to erasure—the right to be forgotten—Article 17(3)(d) of the GDPR caters for a restriction to it where the processing of personal data takes place for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes in accordance with Article 89(1) of the GDPR in so far, as the right to be forgotten is likely to render impossible or seriously impair the achievement of the objectives of the processing.

2.3 Background

The General Data Protection Regulation (EU) 2016/679 offers a digital environment for companies and organizations where they can better trace, secure, and handle data within the IT infrastructure and beyond. In the same vein, the GDPR requires strong security mechanisms to be in place in order to safeguard the data under consideration. Hence, powerful security mechanisms should be adopted for the adequate protection of sensitive and private data stored in order to comply with the GDPR. The latest trends in cybersecurity have embedded technologies with enhanced mechanisms for better results, including machine learning and big data analytics on network security solutions (Kantarcioglu & Xi, 2016).

Chapter 3: A compliant and secure IT infrastructure for the National Library of Greece in consideration of internet security and GDPR

3.1 Overview

The application of the General Data Protection Regulation (GDPR) is considered an issue of vital importance for the smooth operation of IT infrastructures, especially for companies in the European Union (EU) Member States. The GDPR is a useful tool, which, among other requirements, mandates the adoption of privacy-by-design and advanced security techniques. Taking into account its requirements, this article analyzes its implementation with regard to applied Internet Security solutions. While the Regulation offers a minimum set of technical Internet Security means to be taken into consideration by companies and organizations, so as to achieve GDPR compliance, the current chapter aims to highlight the adaptation of strong security mechanisms that will not only set companies compliant with the GDPR but also maintain them strong and secure against most threats.

3.2 Enhanced Security Mechanisms for the National Library of Greece

The General Data Protection Regulation 2016/679/EU (hereinafter, GDPR) offers a digital environment for companies and organizations where they can better trace, secure, and handle data within the IT infrastructure and beyond. In a similar vein, the GDPR requires strong security mechanisms to be in place in order to safeguard the data under consideration. All public and private libraries, including of course the National Library of Greece (hereinafter NLG), need to comply with GDPR requirements for personal data protection. Hence, powerful security mechanisms

should be adopted for the adequate protection of personal data and/or special categories data stored and for compliance with the GDPR requirements. The latest trends in cybersecurity have embedded technologies with enhanced mechanisms for better results, including machine learning and big data analytics on network security solutions (Kantarcioglu & Xi, 2016). In this chapter, we analyze the basic components that NLG should implement and properly configure, in order to become compliant with the relevant legislation and resilient to cyberattacks.

Under GDPR, following a data breach the data controller has the legal obligation to notify the supervisory authority in 72 hours maximum[60]. This way, apart from any actual data losses, organizations run the risk of harming their reputation and even face the financial burden of GDPR fines. Due to the high costs of data breaches (Ponemon Institute LLC, 2017; Ponemon Institute LLC, 2019), security-by-design is highly recommended to companies and organizations to assist them in minimizing investments for their IT security infrastructure and protecting their data. NLG is not an exception to the rule of this recommendation, of course. Strong security mechanisms should be implemented in order to not only be compliant with the GDPR requirements but also be secure against the continuously evolving cybersecurity threats. Once these strong security mechanisms are in place, the level of security in an IT infrastructure will be enhanced and the company will minimize the required response to a security breach.

Thus, NLG has no alternative but to maximize its posture about Information Technology security through technical measures with the aim to enhance the security of its IT infrastructure and comply with GDPR requirement for processing of data in a manner that it *"ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures*[61]*."*

[60] Art.33(1) of the GDPR, according to which "*In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay."*

[61] Art.5(1)(f) of the GDPR.

By the term '*technical measures'* mentioned in the Regulation 2016/679/EU, the legislator refers to the functions, processes, controls, systems, procedures, and policies that are in place to protect and safeguard the critical data and private information that a company holds.

The optimal way to begin is to conduct vulnerability scans and penetration tests on the network and its components, including servers, routers, switches, and endpoints. Risk assessments will assist further to discover loopholes in all processing activities, to identify the highest risks regarding personal data and the effective measures that should be taken under consideration.

Regardless of the possible future changes in the structure of NLG, a resilient IT security infrastructure for NLG should include a firewall solution with endpoint security, a Data Leakage Prevention solution (DLP) which will include a Mobile Device Management (MDM), an Intrusion Detection - Intrusion Prevention System (IDPS), encryption of email communications, encryption and pseudonymization of personal and sensitive data stored in the IT infrastructure. Regarding the authentication and authorization of the users in a web-based service, a strong access control solution should be adopted and included in the arsenal of minimum IT security mechanisms, as well.

Figure 1 depicts a suggested topology regarding a compliant and secure IT infrastructure for NLG in consideration of its structure currently.

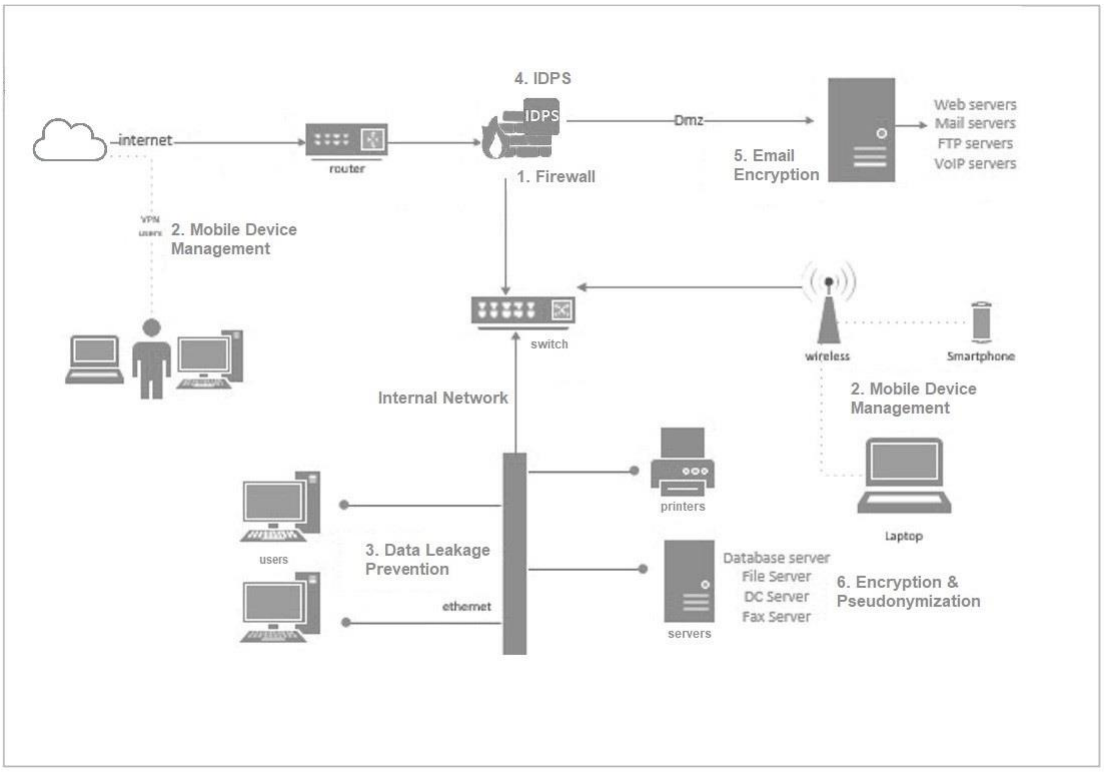


Figure 1: Suggested Topology Regarding a Compliant and Secure IT Infrastructure of NLG.

3.2.1 Firewall

A firewall solution for NLG is an important component, able to safeguard its IT infrastructure, which not only monitors and manages all incoming and outgoing network traffic but, additionally, is able to create a bulletproof environment against all types of malware threats. Numerous attacks have been seen in the wild that mainly have to do with poor configurations and lack of targeted rules enforced (Wool, 2010). While cyber intruders and malicious programs can bypass firewall solutions, these are the primary defense line against cyber-attacks coming from the outside of an IT infrastructure, as well as from insiders or malware attempting to transmit stolen data from the network's interior. Therefore, NLG's firewall is an essential ingredient of compliance with the GDPR and other regulations.

A firewall can be implemented as hardware and software, or a combination of both according to the needs of the IT infrastructure. Typically, a firewall protects against malicious users and allows only legitimate users to access the network, based on the security strictly defined by the IT administrator and the applicable organization's policies. NLG's well-configured firewall solution can determine potential malicious activities, monitor and alert on denial of service and distributed denial of service attacks when occurred, control access to all connected to its network computers if endpoint protection is enabled, and allow access to information based on certain levels of trustworthiness that have been set by NLG's IT administrator and applicable IT security policies.

3.2.2 Data Leakage Detection and Prevention Systems

The GDPR has introduced the concepts of '*Privacy by Design*'[62] and '*Privacy by Default*'[63, 64], which have been topics related to data protection frequently

[62] See Art.25(1) of the GDPR.

[63] See Art.25(2) of the GDPR.

[64] Art.25 of the GDPR, which is titled '*Data Protection by design and by default*'. According to Art.25 of the GDPR "*Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of*

discussed (Ved, 2017). The first thoughts of '*Privacy by Design'* were expressed in the 1970s and were incorporated in the 1990s into Data Protection Directive, i.e. Directive 95/46/EC[65]. According to Recital 46 in Data Protection Directive, Technical and Organizational Measures (TOM)[66] must be taken already at the time of planning a processing system to protect data safety. The term '*Privacy by Design'* means nothing more than data protection through technology design[67]. Behind this, there is the thought that data protection in data processing procedures is best adhered to when it is already integrated into the technology when created. The essence of Article 25 of the GDPR is to impose a qualified duty on controllers to put in place technical and organizational measures that are designed to implement effectively the data protection principles of the GDPR and to integrate necessary safeguards into the processing of personal data so that the processing will meet its requirements and otherwise ensure the protection of data subject's rights.

---

the processing itself, implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article."

[65] See Recital 46 of Data Protection Directive according to which "*Whereas the protection of the rights and freedoms of data subjects with regard to the processing of personal data requires that appropriate technical and organizational measures be taken, both at the time of the design of the processing system and at the time of the processing itself, particularly in order to maintain security and thereby to prevent any unauthorized processing; whereas it is incumbent on the Member States to ensure that controllers comply with these measures; whereas these measures must ensure an appropriate level of security, taking into account the state of the art and the costs of their implementation in relation to the risks inherent in the processing and the nature of the data to be protected."*

[66] See Recital 78 of the GDPR titled *'Appropriate Technical and Organizational Measures*'. The definition of TOMs in the GDPR includes indicatively internal policies and measures which meet in particular the principles of data protection by design and data protection by default, and could consist, inter alia, of minimizing the processing of personal data, pseudonymizing personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features, etc.

[67] See Recital 78 of the GDPR according to which "*When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations."*

The wording of Article 25 of the GDPR expressly describes a duty for data-protection that applies not just at the time of processing but also beforehand, when the controller determines the means for processing, i.e. at the time of designing an information system.

Nevertheless, there is still uncertainty about what '*Privacy by Design*' means, and how one can implement it. On the one hand, this is due to incomplete implementation of the Data Protection Directive in some Member States. On the other hand, the principle of '*Privacy by Design*' which is addressed in the GDPR, requires persons responsible already to include definitions of the means for processing TOMs at the time that they are defined in order to fulfill the '*Privacy by Design*.' The legislation leaves completely open the exact protective measures to be taken. GDPR allows for the definition of TOMs through codes of conduct prepared by library industry bodies[68] or by proper certification schemes[69]. The '*Privacy by Design*' requirement is met through TOMs, i.e. through not just technical measures, but also through organizational measures. In other words, they embrace not simply the design and operation of software or hardware, but they also extend to business strategies and other organizational measures and practices such as rules determining which and under what circumstances NLG employees are authorized to access or otherwise process personal data or special categories data.

The '*Privacy by Design'* and '*Privacy by Default*' mandates of Article 25 of the GDPR is addressed to controllers. Thus, NLG acting as a controller must adhere to the requirements set by Article 25 of the GDPR. However, these requirements must also be met by third parties that NLG leverages upon for the provision of its services in consideration of Recital 78 of GDPR according to which "*When developing, designing, selecting and using applications, services, and products that are based on the processing of personal data or process personal data to fulfill*

[68] See Art.40(2)(h) of the GDPR, according to which "*Associations and other bodies representing categories of controllers or processors may prepare codes of conduct, or amend or extend such codes, for the purpose of specifying the application of this Regulation such as […]"*

[69] See Art.25(3) of the GDPR, according to which "*An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article."*

their task, producers of the products, services, and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfill their data protection obligations". '*Privacy by Design*' is an obligation set by the GDPR that affects also NLG's processors since NLG is only permitted to use processors that provide "*sufficient guarantees to implement appropriate technical and organizational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject*[70]."

In addition to the named criteria, the type, scope, circumstances, and purpose of the processing must be considered. This must be contrasted with the various probability of occurrence and the severity of the risks connected to the processing. The text of the law leads one to conclude that often several protective measures must be combined to satisfy statutory requirements. In practice, this consideration is already performed in an early development phase, when setting technology decisions[71]. Recognized certification can serve as an indicator for the authorities that NLG has complied with the statutory requirements of '*Privacy by Design*[72].'

The outcome of the concepts of '*Privacy by Design'* and '*Privacy by Default'* described in the GDPR is that NLG is now legally accountable for any loss or unauthorized access and usage of the personal data it processes. NLG could further develop its understanding of '*Privacy by Design*' and '*Privacy by Default*' requirement set by Article 25 of the GDPR by delving into the requirements described in the Cybersecurity Act: Regulation 2019/881/EU of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology

[70] See Art.28(1) of the GDPR; see, also, Recital 81 of the GDPR.

[71] See Recital 78 of the GDPR.

[72] See Art.25(3) of the GDPR.

cybersecurity certification and repealing Regulation (EU) Nº 526/2013 (Cybersecurity Act) (Text with EEA relevance)[73].

NLG's Data Leakage Detection and Prevention Systems solution can monitor, restrict, and block the transferring of personal data. NLG's Data Leakage Detection and Prevention Systems solution provide information regarding the organization's data giving the ability to network administrators to enforce rules according to the defined level of sensitivity of the data in question. By adopting these solutions, NLG's network administrator can mitigate data leaks coming from users' errors or internal malicious activities. NLG's Data Leakage Detection and Prevention Systems enhance NLG's GDPR compliance as it is leveraged in order to find, follow, delete, restrict access, prevent access and maintain personal data within NLG's IT infrastructure.

3.2.3 Special Categories Data Storage

The GDPR requires data processors and data controllers to define where personal data and information is stored or processed[74]. By implementing NLG's Data Leakage Detection and Prevention Systems solution, the NLG network administrator is able to scan the existing network infrastructure, including mobile devices and endpoints, for special categories data as defined by policies, compliance profiles, personally identifiable information, file extensions, and other attributes that specify the nature of the data. By doing so, NLG can be informed of the location of special categories data and where they go throughout its network infrastructure. Furthermore, the application of NLG's Data Leakage Detection and Prevention Systems solution makes it is easier for the network administrator to provide extensive reports when requested by the Hellenic Data Protection Agency.

[73] See Regulation 2019/881/EU available at URL: https://eur-lex.europa.eu/eli/reg/2019/881/oj (last check, April 30, 2020). According to Recital 41 of this Cybersecurity Regulation "*ENISA should play a central role in accelerating end-user awareness of the security of devices and the secure use of services, and should promote security-by-design and privacy-by-design at Union level. In pursuing that objective, ENISA should make use of available best practices and experience, especially the best practices and experience of academic institutions and IT security researchers."*

[74] Art.4 No.2 of the GDPR.

3.2.4 Deletion of Special Categories Data

One of the GDPR's requirements is to collect data strictly for the necessary for the pre-defined specific, explicit, and legitimate purposes and not process further the data in a manner that is incompatible with those purposes[75]. Data must be stored only for the limited time that is needed—the principle of data storage limitation (European Commission, 2018). In the framework of NLG's statutory goals and scope of activities, further processing of data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes, in accordance with Article 89(1) of the GDPR[76].

Data must be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. For NLG, the storage limitation principle of the GDPR allows for data storage for longer periods "*insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to the implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject*[77]."

By implementing NLG's Data Leakage Detection and Prevention Systems solution, the NLG's network administrator becomes able to delete personal data or special categories data stored anywhere in the NLG's network infrastructure even remotely, allowing for total control over the stored data within the NLG's IT infrastructure, and at the same time making possible instant decisions on the maintenance of data on any devices connected to NLG's network ("3 Phases of protection by a Data Leakage Prevention (DLP) plan", 2017).

[75] Art.5(1)(b) of the GDPR.

[76] Art.5(1)(b) of the GDPR.

[77] Art.5(1)(e) of the GDPR.

3.2.5 Restriction of Special Categories Data usage

Another GDPR requirement is to ensure that the controller or the processor will not process special categories data for any other purposes besides those strictly referred to.[78] Processing of special categories data by the NLG is allowed provided that: said "*processing is necessary for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject*[79]*."*

NLG does not upload or share data of special categories to any cloud service. By applying NLG's Data Leakage Detection and Prevention Systems solution, special categories of data can be identified by the NLG's network administrator, who restricts and totally denies any kind of data transfer within the network infrastructure or outside of it, by using rules, policies, and specific filters. Consequently, users of NLG's connected computers are restricted from uploading, downloading, copying or printing, or otherwise using any special categories data.

3.2.6 Maintaining Security Standards

Following the enforcement of the GDPR, NLG as a data controller is required to be informed regarding the privacy and security standards that processors, which NLG may leverage upon, have adopted and implemented and whether or not they have been up to date[80]. Through the implementation of NLG's Data Leakage Detection and Prevention Systems solution, transferred or stored data can be scanned within the controller and processors' IT infrastructure. Thus, if a data breach occurred, NLG can inform the processors in order for the latter to take

[78] Art.9(1) of the GDPR.

[79] Art.9(2)(j) of the GDPR.

[80] Art.32 of the GDPR.

specific actions. NLG's Data Leakage Preventions system is an integral part of GDPR implementation by preventing losses of personal data and/or special categories data in NLG's network infrastructure, covering two of the most important components of GDPR, the integrity and confidentiality of protected data[81] (European Data Protection Supervisor, 2018).

3.2.7 Mobile Device Management system

Mobile device management (MDM) is a type of security software that can be used by the NLG's IT department to monitor, manage and secure employees' mobile devices that are deployed across multiple mobile service providers and operating systems used in the organization (Blokdyk, 2019). The GDPR has a strong impact on the way that NLG can handle its data within mobile devices.

A great change that came with the GDPR, is the need-to-know basis of where the data exist at any given moment and the consent of the individual for storing and using the data. Another requirement is to know the origin of specific data and the individual who shared these data, which is a challenge especially for mobile users, for example, NLG's personnel who may collect data through its different premises.

One of the most important functionalities in order to both meet a certain security level and be compliant with the GDPR is to determine the devices that have access to specific data and services at all times. NLG's Mobile Device Management solution allows to include some personal devices which need to be separate objects of a risk assessment before allowing their users to include protected data. NLGs' Mobile Device Management solution should come part and parcel with NLG's Bring Your Own Device policy that strictly describes the nature of the data and the access level through mobile devices policy applied in NLG (BYOD) (French et al., 2014). Gathering all the information and devices included in NLG's Mobile Device Management solution and Bring Your Own Device

[81] Art.5(1)(f) of the GDPR.

policy, the NLG's IT administrator will have the ability to audit the logs and specify the actions that took place in an event of data breach.

It is of high importance to pinpoint that, more often than not, mobile devices are overlooked when it comes to security. However, they constitute a great risk when it comes to security and compliance if the appropriate security mechanisms are not in place. In any case, it is important to adopt strict security controls including proper configuration, policies, and encryption techniques on every device, in order to safeguard the data included.

Of course, ensuring the security of business data is much easier when personal and business data are kept separate. Establishing clear boundaries between a user's personal data and NLG's business data on personnel's mobile devices is an important, although hard step to take. Ideally, any user including NLG's personnel should not be able to gain access to any personal apps or personal email accounts on a business device and vice versa. This would help minimize security risks and it is a way for NLG to stay GDPR compliant regarding the integrity and confidentiality of protected data. However, this is not an easy step to take, so NLG needs to focus on minimizing the overlap and establishing clear boundaries and policies for managing TOMs deployed for securing the integrity and confidentiality of protected data.

3.2.8 Intrusion Detection and Prevention system

Cyberattacks are still one of the biggest reasons for personal data compromising. Verizon's "2017 Data Breach Investigations Report", mentions that more than half of the breaches occurred (51%) can be traced back to malware, and this is just one type of network intrusion that can lead to data being compromised (Ward & Pritam, 2017).

An Intrusion Detection and Prevention system is one of the most powerful ways to get properly protected from cyber threats acting in a different way from the other security components mentioned so far in this chapter. NLG's firewall for

example is able to filter potentially malicious traffic coming into NLG's network, while NLG's Intrusion Detection and Prevention system is able to monitor all traffic inside NLG's network infrastructure and alert the NLG IT administrator if malicious acts have been detected.

NLG's Intrusion Detection and Prevention system is a device or software application that monitors NLG's network or systems for malicious activity or policy violations. Any malicious activity or violation is typically reported either to the NLG's administrator or is collected centrally using a Security Information and Event Management (SIEM) system. A SIEM system combines outputs from multiple sources and uses alarm-filtering techniques to distinguish malicious activity from false alarms (Blokdyk, 2020).

As it is shown in the network topology depicted above and according to the needs of the NLG's IT infrastructure, NLG's Intrusion Detection and Prevention system should be placed after the firewall, so only legitimate traffic is inspected, which will further reduce the load on it as well. However, there are IT infrastructures where it is advisable to place the Intrusion Detection and Prevention system in front of the firewall to protect the firewall from cyberattacks. Therefore, a strong Intrusion Detection and Prevention system solution customized to the needs of NLG will enhance the protection from cyberattacks.

3.2.9 Email encryption

Sending and receiving emails are considered high-risk activities regarding information security because it is possible for a network administrator, the service provider, or even a malicious user to capture this kind of communication. As a result, it is of high importance to use encryption techniques when transferring sensitive data via email communications and strongly recommended for GDPR compliance and overall security[82]. Even though email encryption is considered an

[82] Art.32 of the GDPR.

important component for security in general, it is not commonly implemented within the IT security infrastructures for companies and organizations.

There are two basic techniques for email encryption, each one serving on a different level. The first method encrypts the emails while transmitted from one end to the other via an encrypted tunnel. Emails are encrypted at the source before the actual transfer, and decrypted upon arrival to the destination, using network protocols such as TLS (Transport Layer Security) and its ancestor SSL (Secure Socket Layer). The second method for email encryption encrypts the content of the email without interfering with any transferring protocols while transporting the email. Thus, even if the packets are captured by anyone in the middle of the communication, the content cannot be shown as it is encrypted (Desmedt, 2005).

A known and widely used method regarding content encryption is Secure/Multipurpose Internet Mail Extensions (S/MIME) and Open PGP based on Pretty Good Privacy (PGP) (Callas et al., 2007). PGP is an encryption program that provides cryptographic privacy and authentication for data communication (Zimmermannn, 1995). Regarding email encryption nowadays, there are numerous solutions by enterprise vendors like Microsoft (Microsoft Corp., 2019) and Symantec (Symantec Corp., 2019).

3.2.10 Encryption

In a similar vein with email encryption, data encryption converts clear text into a hashed code using ciphers, where the encrypted information transforms to readable again by the method of decryption. It needs numerous resources to decrypt a ciphertext without the knowledge of the encryption key, thus, it is highly unlikely for a third party to decrypt a captured ciphertext. Strong encryption is the best solution for NLG in order to safeguard the transmission of sensitive information and one of the best methods to protect personal data or special categories data by unauthorized access while stored.

3.2.11 Pseudonymization

Pseudonymization is a data-management and de-identification procedure by which personally identifying information fields within a data record are replaced by one or more artificial identifiers or pseudonyms. A single pseudonym for each replaced field or collection of replaced fields makes the data record less identifiable while remaining suitable for data analysis and data processing (International Organization for Standardization/Techincal Committee, 2008). Regarding compliance to the GDPR, the application of pseudonymization to personal data can reduce the risks for the data subjects concerned and help controllers and processors to meet their data-protection obligations[83]. Pseudonymization is recommended as a means for securing data protection in the NLG's network and IT infrastructure but should not be used as a way to separate identifiers from data subjects regarding personally identifiable information in order to circumvent other obligations ("Personal data pseudonymization: GDPR pseudonymization what and how", n/d).

3.2.12 Non-Technical Measures

Apart from strong technical implementations, NLG needs an effective Information Security Policy to be adopted as an essential part of its IT-security posture. This policy could include subsection such as Asset Management, Access Control, Passwords & Encryptions, Remote Access, Bring Your Own Device (BYOD), Clear Desk & Screen, Secure Disposal, Business Continuity Plan/Disaster Recovery. In addition, security awareness and training for all the NLG employees are required in order to avoid security breaches. Cybercriminals will mostly target the weakest link in the chain of security, which is the human factor (Evans, 2009). An employee unintentionally could click on a phishing email, share sensitive information over the telephone, and could let ransomware be installed in the computer connected to the NLG's network, which could cause serious damage to the organization (Scheeres, 2012). For this reason, the NLG's employees should

[83] See Recital 28 of the GDPR titled *Introduction of Pseudonymization*.

become aware and be periodically trained on best practices in the usage of their systems and their obligations and responsibilities regarding the data they manage and are associated to. Training on a regular basis regarding strong password policies, best practices while surfing online, locking the screen when leaving the desk, and the latest techniques used by hackers in order to evade a system, are some of the recommended practices that promote effectively throughout the NLG secure IT infrastructure and work as an add-on to a secure by design IT infrastructure within the organization.

3.2.13 Authentication and Authorization on a web service

TDM has already been in use for security threat detection, and for discovering hidden information in unstructured log messages (Suh-Lee et al., 2017). Moreover, TDM has been used for security and crime detection (Paaβ et al., 2014). While TDM may be used for security purposes, the NLG's databases that contain and manage sensitive, confidential, and valuable data, should be properly protected from unauthorized access and data losses. Successful implementation of data security at NLG should not only provide accurate and timely data but also protect its confidentiality, integrity, availability, and security overall. While creating a database that is accessible online by registered users such as NLG's general catalog[84], the NLG's administrator must properly safeguard the catalog's stored data. NLG's general catalog and other NLG services available through the Web appear more vulnerable than other NLG infrastructures because they are accessible online, thus anyone can potentially access them. Consequently, the addition of a new set of requirements to the security landscape for all NLG services and the NLG's Web services is a necessity. NLG's properly configured Access Control policy is one of the most important components regarding security on Web services offered from the library and it works as a powerful tool to protect the

[84] See NLG's catalogue available to registered users at https://www.nlg.gr/collection/catalogue/ (last check, April 30, 2020).

stored data accessed by external users, as it filters who will have access and where exactly.

One of the major aspects regarding the security of the NLG Web services is the authentication of each user trying to access the web service under consideration, verifying that the user is who claims to be. Each user is identified through the NLG's Single Sign-On system by providing valid credentials, mostly their full name, email address, and date of birth[85]. Users of the NLG Web services are authenticated through SSO which is connected to Independent Authority for Public Revenue's system[86].

3.3 Conclusions

Throughout this chapter, we focused on NLG's Internet Security solutions in consideration of compliance with the GDPR. While the Regulation offers a minimum set of technical Internet Security means to be taken into account by companies and organizations with the aim to achieve GDPR compliance, the current chapter highlights a set of TOMs deemed necessary for the application of strong security mechanisms at NLG. By adopting the TOMs described in this chapter, NLG can enhance its ability to better protect the library's IT infrastructure from cyber threats. It can also improve the response to these kinds of threats and mitigate their impact. Even after more than two years since the GDPR has been enforced, a large number of companies including most public libraries have not adopted a strong security strategy rendering them vulnerable to cybersecurity threats. The GDPR is an opportunity for cybersecurity in practice, giving the chance to companies and organizations to implement effective security mechanisms. Apart from the fines enforced by Regulation 2016/679/EU, protecting

[85] See NLG's SSO system available at https://register.nlg.gr (last check, April 30, 2020).

[86] See login through NLG's SSO system available at https://sso.nlg.gr/login (last check, April 30, 2020).

personal data and data of special categories is foremost a matter of the NLG's IT infrastructure security.

NLG, as well as all national libraries of the other EU Member States, are organizations which operate as safeguards of cultural treasures: they aim to build a distributed and permanent collection of digital resources from the field of digital preservation development of a distributed network of safe-kept material with resource owners, or parties nominated by them, providing long-term access to their material. For this reason, Directive 2019/790/EU on copyright and related rights in the Digital Single Market, which amends Directives 96/9/EC and 2001/29/EC, paves the way for NLG and other EU Member States' national libraries to proceed with massive preservation of cultural heritage. Copyright law that negatively impacted on the reproduction of works protected by copyright is being amended through Article 6 of Directive 2019/790/EU which caters for a mandatory exception "*to the rights provided for Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) of Directive 2009/24/EC and Article 15(1) of this Directive, in order to allow cultural heritage institutions to make copies of any works or other subject matter that are permanently in their collections, in any format or medium, for purposes of preservation of such works or other subject matter and to the extent necessary for such preservation*." The digitization, maintenance, connectivity, and the making available to the public of digital resources kept in national libraries are becoming core features in an ever-growing number of products and services offered by them and with the advent of the Internet of Things (IoT), a high number of connected digital devices are expected to be deployed by national libraries across the EU during the next decade. While an increasing number of cultural resources and treasures safe-kept in NLG and other national libraries are becoming available online, and while an increasing number of devices connected to the internet may be leveraged upon to access these cultural resources, security and resilience are not sufficiently built-in by design in national libraries, leading to insufficient cybersecurity.

Cybersecurity is not only an issue related to technology but one where human behavior is equally important. The so-called, '*cyber hygiene*[87],' namely, simple, routine measures that minimize their exposure to risks from cyber threats, when implemented and carried out regularly by citizens, organizations, and businesses should be strongly promoted in the environment of NLG and all national libraries of EU Member-States. National libraries, which most—if not all—of them are public organizations, are involved in the design and development of ICT products, ICT services, or ICT processes, thus they should be encouraged to implement TOMs at the earliest stages of design and development to protect the security of those products, services, and processes to the highest possible degree, in such a way that the occurrence of cyberattacks is presumed and their impact is anticipated and minimized ('*security-by-design*')[88]. Security should be ensured throughout the lifetime of an ICT product, ICT service, or ICT process deployed in a national library by design and development processes that constantly evolve to reduce the risk of harm from malicious exploitation. National libraries' ICT products or ICT services or ICT processes should be designed and be made available in such a way so that they ensure a higher level of security which "*should enable the first user to receive a default configuration with the most secure settings possible* ('*security by default*'), *thereby reducing the burden on users of having to configure an ICT product, ICT service or ICT process appropriately*[89]."

All EU Member States' national libraries including NLG, of course, have the daunting task of compliance with the GDPR. This task can be seen in the wider spectrum of creating cyber-resilient organizations, i.e. national libraries that consider compliance with the requirements of Directive 2016/1148/EU concerning measures for a high common level of security of network and information systems across the EU. Most of the cyber-security issues are common to all national

[87] See Recital 9 of Regulation 2019/881/EU on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act).

[88] See Recital 12 of Regulation 2019/881/EU.

[89] See Recital 13 of Regulation 2019/881/EU.

libraries in the EU. The establishment of a European cybersecurity certification framework through Regulation 2019/881/EU[90] that lays down the main horizontal requirements for European cybersecurity certification schemes to be developed and allows them alongside EU statements of conformity for ICT products, ICT services or ICT processes to be recognized and used in all Member States, could be leveraged in the industry of EU national libraries through a cybersecurity certification scheme customized to the relevant needs and requirements of national libraries, museums, and archives in the EU. We refer to a European certification scheme laying down the comprehensive set of rules, technical requirements, standards, and procedures that are established at the Union level and that apply to the certification or conformity assessment of specific ICT products, ICT services, or ICT processes used by libraries, museums, and archives throughout EU. The idea for a European cybersecurity certificate addressing the cybersecurity needs of libraries, museums, and archiving organizations Europe-wide could be cultivated and further developed through organizations such as the European Bureau of Library, Information, and Documentation Associations (ELBIDA) or organizations with international poise such as the International Federation of Library Associations and Institutions (IFLA) which could furnish the European Cybersecurity Certification Group provisioned in Article 62 of Regulation 2019/881/EU with information on the special needs and requirements for secure ICT products, ICT services, and ICT processes aimed to be used by organizations such as national libraries in the EU. A European cybersecurity certification for national libraries, museums, and archiving organizations could be leveraged upon to prove compliance with the GDPR requirements, in consideration of the provision of Article 54(4) of Regulation 2019/881/EU according to which "[…] *a European cybersecurity certification scheme may be used for establishing the presumption of conformity with legal requirements*."

[90] See Art.46 and Art.49 of Regulation 2019/881/EU.

Chapter 4: Compliant and secure websites for the Greek Libraries Network of the National Library of Greece and each library-member of this Network in consideration of internet security and GDPR

4.1 Overview

The application of the General Data Protection Regulation (GDPR) regarding the operation of websites is considered of vital importance, especially to organizations within the European Union. The GDPR is a useful tool, which, among other requirements, mandates the adoption of privacy-by-design and advanced IT security mechanisms in place. Considering its requirements, this chapter analyzes its implementation with regard to applied Internet Security solutions for the websites of the Greek Libraries Network of the National Library of Greece. While the GDPR offers a minimum set of technical Internet Security means to be taken into consideration by companies and organizations Europe-wide, hereby we aim to highlight the adaptation of strong and proper security mechanisms that will not only set libraries-members of the Greek Libraries Network of the National Library of Greece compliant with the GDPR but also maintain them strong and secure against most threats targeting websites to both internal and external cybersecurity threats.

4.2 Security by design and by default for the Greek Libraries Network of NLG

The General Data Protection Regulation 2016/679/EU (hereinafter, GDPR) offers a digital environment for companies and organizations where they can better trace, secure, and handle data within the IT infrastructure and beyond. In this context, the GDPR requires strong security mechanisms to be in place in order to safeguard the data under consideration. All libraries, public and private, need to comply with the GDPR

requirements for personal data protection. Hence, powerful security mechanisms should be adopted for the adequate protection of personal data and/or special categories of data processed by the IT infrastructure. The latest trends in cybersecurity have embedded technologies with enhanced mechanisms for better results, including machine learning and big data analytics on network security solutions (Kantarcioglu et al., 2016). In this chapter, we analyze the basic components that the Greek Libraries Network of the National Library of Greece (hereinafter, NLG Network) supports and coordinates, and the components it should implement and properly configure, so as its websites become compliant with the relevant legislation and resilient to cyber-attacks.

The NLG Network has been established to help Academic, Research, Public, Municipal, and School libraries to develop and evolve the services they offer to their public. It aspires to be the means of exchanging information, knowledge, and professional communication between libraries, to undertake training program initiatives for library staff, to plan and organize national and international activities, such as campaigns or conferences, to inspire its members to the use of professional tools and standards, and to provide ongoing support to libraries which are members of the Network. It provides leadership and consultant services to all its member (e.g. applying standards and implementing new). In the organizational field, the NLG Network offers library support by creating a union catalog, providing Integrated Library Systems (ILS), and defending the political, economic, and social positions of libraries so that in the future they can be social hubs for the community they serve. It aims to conduct research and surveys to serve its purpose (e.g. level of development of library services in the country, benchmarking based on other countries' data and metrics). Also, the NLG Network has a key role in the preparation of integrated funded programs, either national or international, that aim at the development of libraries in the Network.[91] NLG and all member-libraries of the Greek Libraries Network set up their collaboration through the NLG Network with the aim to apply Network-wide innovative methods of developing services and programs that each library could not implement on its own, such as Information Literacy, and jointly and radically comply to the requirements of Library

[91] Dr. Philippos Tsimpoglou, General Director of the National Library of Greece, in Network of the Greek Libraries of the National Library of Greece, available at URL: https://network.nlg.gr/liga-logia/.

Science in the modern landscape in Greece today, which of course includes the application of the GDPR requirements in the operation of libraries. For all the above reasons, the Greek Libraries Network has the pivotal role of mapping and organizing the Greek libraries' eco-system and establishing an annual survey documenting resources, services, and developments.

Under the GDPR, following a data breach the data controller has the legal obligation to notify the supervisory authority in 72 hours maximum[92]. This way, apart from any actual data losses, organizations run the risk of harming their reputation and even face the financial burden of GDPR fines. Due to the high costs of data breaches (Ponemon Institute LLC., 2017; Ponemon Institute LLC., 2019), '*security by design'* and '*security by default'* are highly recommended to for-profit and non-profit organizations, such as most libraries of the NLG Network, in order to be assisted with minimizing investments for their IT security infrastructure and protecting their data. Once these strong security mechanisms are in place to cover a potentially wide range of data protection measures and to ensure data minimization and confidentiality, the level of security of an IT infrastructure, which is used for the processing of personal data, is enhanced and the organization minimizes the required response to a security breach.

Therefore, in order for the NLG Network to enhance the security of its websites and comply with the GDPR requirement for processing of data in a manner "*that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures*[93]*"*, it needs to maximize its Information Technology security posture through proper "*technical or organizational measures*". By the term "*technical or organizational measures*", mentioned in Regulation 2016/679/EU, the legislator refers to the functions, processes, controls, systems,

[92] Art.33(1) of the GDPR, according to which "*In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay."*

[93] Art.5(1)(f) of the GDPR.

procedures, and policies that are in place, to protect and safeguard the critical data and private information that an organization holds.

GDPR's Article 25[94] mandates the requirement for data protection '*by design'* and '*by default'*, leveraging on '*technical and organizational measures'*, taking into account the state of the art, the cost of implementation, and the nature, scope, context, and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing. The controller is mandated to implement, both at the time of the determination of the means for processing and at the time of the processing itself, '*appropriate technical and organizational measure's*, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of the GDPR and protect the rights of data subjects. This obligation ensures by default that only personal data that are necessary for each specific purpose of the processing are processed. This applies to the amount of personal data collected, the extent of their processing, the period of their storage, and their accessibility. *'Appropriate technical and organizational measures'* must ensure by default that personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

Clarification upon the notion of *appropriate technical and organizational measures* is provided in Recital 78 of the GDPR, according to which "*such measures could consist, inter alia, of minimizing the processing of personal data, pseudonymizing personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features".* Technical and organizational measures must be taken *"to ensure that the requirements of this Regulation* [a.k.a. GDPR] *are met";* technical and organizational measures consist of policies and measures which "*meet in particular the principles of data protection by design and data protection by default".* The principles of data protection '*by design'* and '*by default'* must be taken into

[94] See Art.25(1) & (2) of the GDPR.

consideration in the context of public tenders, too.[95] Therefore, any attempt of the NLG Network to apply for integrated funded programs either national or international that aim at the compliance of libraries in the Network to the requirements of the GDPR must consider data protection '*by design'* and '*by default*.'

In consideration of the principles of data protection *'by design'* and *'by default'* "*when developing, designing, selecting, and using applications, services, and products that are based on the processing of personal data or process personal data to fulfill their task, producers of the products, services, and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfill their data protection obligations."*

Recital 46 of the Data Protection Directive 95/46/EC (hereinafter, DPD) mentions the need to take '*appropriate technical and organizational measures'* for the protection of data subjects' rights and freedoms "*both at the time of the design of the processing system and at the time of the processing itself, particularly in order to maintain security and thereby to prevent any unauthorized processing*." Since the application of DPD in personal data protection, the notion of '*appropriate technical and organizational measures'* is aligned with measures that ensure an appropriate level of security, taking into account the state of the art and the costs of their implementation in relation to the risks inherent in the processing and the nature of the data to be protected.

The Court of Justice of the European Union (hereinafter, CJEU) has yet to rule on the subject matter of Article 25 of the GDPR. However, it has strongly ruled through joined cases C-293/12 and C-594/12 that the adoption of '*technical and organizational measures'* ensures that personal data are given effective protection against the risk of abuse and against any unlawful access and use.[96] Thus, CJEU is in line with the application of Article 25 of the GDPR which, by its turn, is in line with the rulings of the

[95] See Recital 78 of the GDPR.

[96] See joined cases C-293/12 and C-594/12 *Digital Rights Ireland and Seitlinger and Others*, Judgement of the Court, paras.5, 7, 40, 66, 67, available at URL: http://curia.europa.eu/juris/document/document.jsf?text=&docid=150642&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=10218830 .

CJEU even before the enactment of the GDPR. The CJEU has also ruled indirectly in line with the requirements of Article 25 of the GDPR in some of its decisions dealing with internet mechanisms, such as the decision in Case C-131/12 in which the CJEU ruled that Google and other search engine operators must reconfigure their systems so that they are more privacy-friendly.[97]

The purpose of Article 25 of the GDPR is to impose a qualified duty on controllers to put in place technical and organizational measures that are designed to implement effectively the data protection principles of the GDPR. Article 25 of the GDPR prevents controllers from using technologies that collect more personal data than are strictly necessary for technological functionality or that leak personal data to outsiders. The wording of the GDPR expressly applies its ruling not just at the time of the processing but also beforehand when the controller determines the means for the processing. The measures referred to in Article 25 of the GDPR are not just '*technical'* but also '*organizational*,' which means that they embrace not simply the design and operation of software and hardware, but they extend to business strategies and other organizational measures which contribute to the application of the GDPR principles listed in Article 5 of the GDPR regarding data processing and privacy protection (Kuner et al., 2020).

Article 25 addresses the notions of '*privacy by design'* and '*privacy by default'*, as well. Both '*privacy by design'* and '*privacy by default'* are to be taken by controllers, which are the entities to determine the purposes and the means of the processing of personal data.[98] '*Privacy by design'* is addressed in Article 25(1) of the GDPR while '*privacy by default'* is addressed in Article 25(2) of the GDPR. Article 25(1) of the GDPR formulates the design stage in terms of when the controller assumes its status by the time of determination of the means for processing.

[97] See Case C-131/12, *Google Spain and Google*, available at URL: http://curia.europa.eu/juris/liste.jsf?oqp=&for=&mat=or&lgrec=el&jge=&td=%3BALL&jur=C%2CT%2CF&num=C-131%252F12&page=1&dates=&pcs=Oor&lg=&pro=&nat=or&cit=none%252CC%252CCJ%252CR%252C2008E%252C%252C%252C%252C%252C%252C%252C%252C%252Ctrue%252Cfalse%252Cfalse&language=en&avg=&cid=10219757.

[98] See Art.4(7) of the GDPR.

Both '*privacy by design'* and '*privacy by default'* impose on library members of the NLG Network the obligation to meticulously select '*technical and organizational'* measures that are privacy-friendly and can ensure effective protection against the risk of abuse and against any unlawful access and use of personal data. Having that in mind, where does this selection start?

One way to begin would be to conduct vulnerability scans and penetration tests on the website and its components, including hosting servers and service providers checks. Endpoint risk assessments could assist further to discover loopholes in all processing activities, identify high risks regarding personal data and effective reparative measures.

Regardless of possible future changes in the IT support system of the NLG Network, a resilient IT security infrastructure for the websites of the Network members should consider the security and regulatory requirements described in the current article for coping with the most common website cybersecurity threats, analyzed in the following section.

4.3 Common website Cyber Security Threats

In order to understand the needs of an organization and conduct the websites of the NLG Network compliant and resilient to cybersecurity threats, we need to thoroughly understand the online threats that might affect the web infrastructure of an organization, including the website. As described on the OWASP Top Ten security risks to web applications of The Open Web Application Security Project[99], the most common threats that must be taken under consideration include, but may not be limited to, the following: Injections, Broken Authentication, Sensitive Data Exposure, XML External Entities (XXE), Broken Access Control, Security Misconfiguration, Cross-Site Scripting XSS, Insecure Deserialization, Using Components with Known Vulnerabilities and Insufficient Logging & Monitoring.

[99] See the OWASP Top Ten about the most critical security risks to web applications of the Open Web Application Security Project at URL: https://owasp.org/www-project-top-ten/.

Regarding these cybersecurity threats:

Injection: Injection flaws are a class of security vulnerability such as SQL, NoSQL, OS, and LDAP injection that allows a user to "*break out*" of the web application context. These vulnerabilities may exist in the case of uncontrolled data entering an interpreter as part of a query or command. In the case of a website or a web application that takes user input into a back-end database, shell command, or operating system call, the application may be susceptible to an injection vulnerability (Bashah et al., 2011).

Broken Authentication: Regarding web applications, the authentication procedure is considered as "*broken*" when a potentially malicious user compromises passwords, keys or session tokens, user data, and further information that may provide more details about the users' identity. Broken authentication attacks aim to achieve unauthorized access to active accounts giving the malicious user the same privileges as the legitimate user by manipulating the improper implementation of application functions related to authentication and session management (Maruf et al., 2018).

Sensitive Data Exposure: This kind of cyber threat may occur in a case of a web application that does not properly secure sensitive data that may include passwords, session tokens, financial data, and even private health data and more. The administrators and developers of web applications should protect properly all sensitive data, including the aforementioned categories, in order to prevent unauthorized access and data losses from malicious users. (Sue et al., 2015)

XML External Entities (XXE): XML external entity injection (hereinafter, XXE) is a flaw targeting web applications, which permits malicious users to manipulate with a web application processing of XML data. By XXE, a potentially malicious user discloses internal files of the web application and interacts with any backend or external systems that the web application itself accesses. Furthermore, an attacker may be able to view internal file shares, conduct internal port scanning, conduct remote code execution, perform Server-Side Request Forgery (SSRF) attacks and perform denial of service attacks (Osincev et al., 2019).

Broken Access Control is another common vulnerability that occurs when testing the security of a web application, which in most cases results in unauthorized access,

information disclosure of sensitive files and data, modification of access rights, and change or corruption of information and data. Broken Access Control is a security flaw that may occur due to a lack of proper security methodology adopted by developers in the procedure of creating a web application during coding or insecure implementation of authentication and authorization mechanisms (Hassan et al., 2018).

Security misconfiguration threats are the most common problems regarding the security of web applications. They occur when a web application does not have proper configuration regarding its source code or the infrastructure that is hosted and may exist in software components or subsystems. It may lead to the exploitation of other vulnerabilities that target any part of the application stack. For the avoidance of any security misconfiguration, it is strongly recommended to harden all the components of a web application including operating systems, frameworks, and libraries, upgrading each part, especially with security patches whenever is needed (Eshete et al., 2011).

Cross-site scripting (hereinafter, XSS) is one of the most common web application security vulnerabilities that is found during website security tests. XSS vulnerabilities provide the ability to a malicious user to inject untrusted content or client-side scripts on web pages without proper data validation techniques or escaping. The manipulation of an XSS vulnerability enables a malicious user to run unauthorized scripts in a victim's web browser leading to the user's session hijacking, website defaces, or redirection to malicious websites. XSS flaws' impact may differ according to the sensitivity of the data handled by the vulnerable application and the nature of any security mitigation implemented by the administrator of the web application (Gupta et al., 2015).

Insecure deserialization is a security flaw where malicious content often leads to remote code execution, denial of service attacks (DoS attack), replay attacks, injection attacks, privilege escalation attacks, authentication bypass, or manipulation of the logic the web application operates (Dehalwar et al., 2017).

The use of application components that include known vulnerabilities is another common security flaw of a web application. It should be underlined that libraries, frameworks, and other software modules have the same permissions as the web application runs. In case of exploitation of a component with known security flaws, this

may lead to further information leakage or even manipulation of the web server where the web application is hosted. Web applications and APIs that include components with known security flaws may be targeted by malicious users and lead to serious security incidents (indicative lists of such security flaws can be found in https://www.exploit-db.com/ and https://attack.mitre.org/).

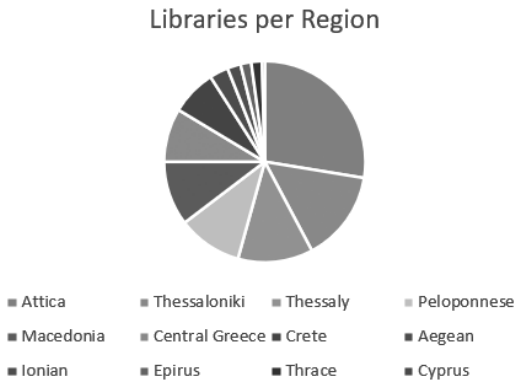
Insufficient Logging & Monitoring: One of the major flaws of a web application is the lack of sufficient logging and monitoring. In the case of an effective security incident, without proper logging and monitoring the web applications' administrator will not be alerted, allowing the malicious user to continue the attack uninterrupted apart from the web application, to the whole infrastructure hosting the application. Typically, for a security incident to be identified, it takes approximately 200 days based on the 2019 Cost of a Data Breach Study sponsored by IBM (Ponemon Institute LLC., 2019) and it is detected by external parties rather than inside procedures or internal security teams (Leite et al., 2019).

4.4 The current situation of library members' websites of the NLG Network

In order to depict the current situation of the websites of library members of the NLG Network, we have analyzed for every member the existence of its website —its social media presence, if any— of an updated Privacy Policy with regard to the GDPR and the use of a Secure Socket Layer (SSL) certificate, regarding the security of the communication between users and the library's website. The NLG Network consists of 234 library members in total, 233 of which are based in Greece and 1 is based in Cyprus.

Table 1. The Greek Libraries Network of the National Library of Greece consists of 234 Libraries. Separation per region.

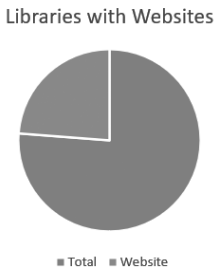
Regions	Libraries per Region
Attica	64
Thessaloniki	34
Thessaly	28
Peloponnese	24
Macedonia	24
Central Greece	20
Crete	17
Aegean	7
Ionian	5
Epirus	4
Thrace	4
Cyprus	1

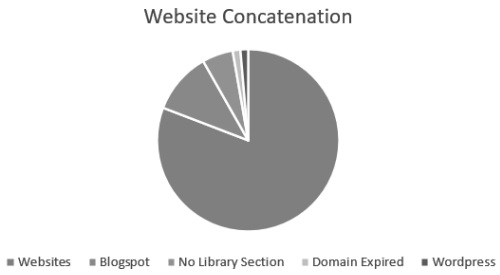


The distribution of the NLG Network as shown in Table 1, includes seven libraries in the region of Aegean, four libraries in the region of Epirus, 28 libraries in the region of Thessaly, 34 libraries in the region of Thessaloniki, 17 libraries in the region of Crete, 24 libraries in the region of Peloponnese, 64 libraries in the region of Attica, four libraries in the region of Thrace, five libraries in the region of the Ionian sea, 20 libraries in Central Greece, one in Cyprus and 26 in the region of Macedonia.

Table 2. The Greek Libraries Network of the National Library of Greece includes 73 libraries with a website.

	Libraries
Total	234
Website	73
	Libraries
Websites	59
Blogspot	8
No Library Section	4
Domain Expired	1
Wordpress	1

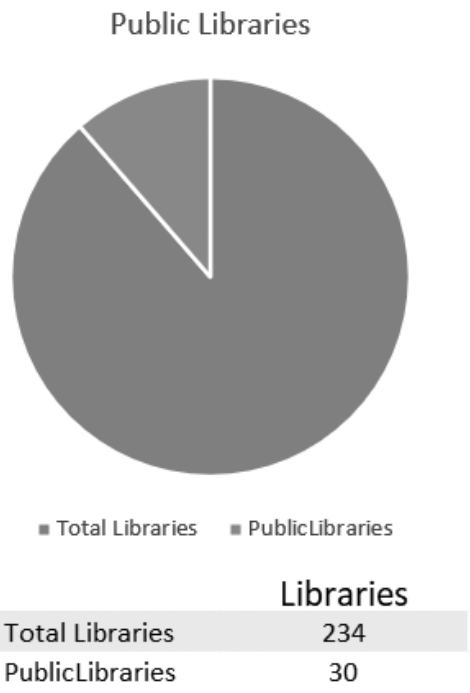




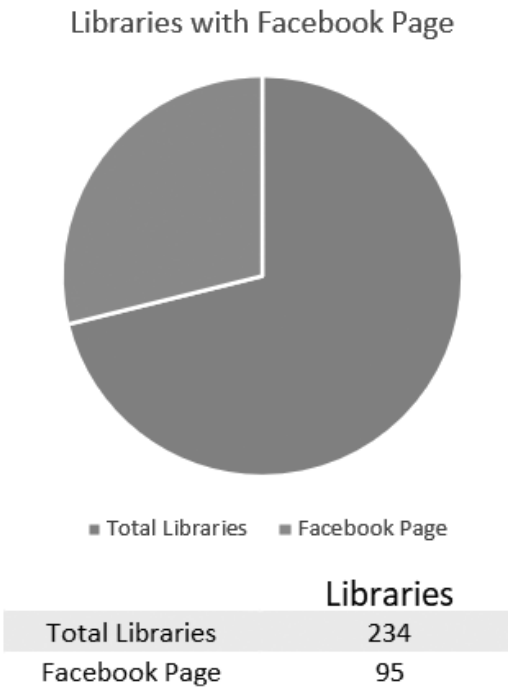
From the total of 234 libraries, 30 are public libraries, 73 have a website and 95 have a Facebook page.

Out of the 73 libraries that have a website, there is either a separate website for the library (e.g. http://vivl-atalant.fth.sch.gr/) or the library is mentioned as a sub-category on a municipal website (e.g. https://mykonos.gr/the-island/culture/). There are also eight libraries that use a BlogSpot page instead of a website, four libraries that are not even mentioned on the municipal website, one library with a WordPress page and one library whose domain has expired.

Table 3. The Greek Libraries Network of the National Library of Greece includes 30 libraries that are Public and 95 have a Facebook Page.



	Libraries
Total Libraries	234
PublicLibraries	30



	Libraries
Total Libraries	234
Facebook Page	95

An important factor regarding website security is the use of Secure Sockets Layer (hereinafter, SSL) certificates. SSL certificates play a crucial role in website security, enabling web applications to adopt the more secure protocol Hypertext Transfer Protocol Secure (hereinafter, HTTPS), which is used for secure communications in the online world, rather than Hypertext Transfer Protocol (HTTP) which is less secure due to the lack of proper encryption. An SSL certificate is a data file hosted on a website's origin server. With the use of SSL certificates, the encryption SSL/TLS occurs. The SSL

certificates include the public key of the website a user's browser is trying to connect to and further relevant information about the website's identity. Users that try to initiate communication with a specific web server reference the exact file containing the above information in order to obtain the public key and verify the server's identity. The private key is not transmitted but is kept secret and secure.

SSL, more commonly called Transport Layer Security (TLS), is a protocol for the encryption of internet traffic and the verification of a server's identity. Any website with an HTTPS web address uses the protocol SSL/TLS. Previous studies have shown that the state of non-browser SSL code is catastrophic across web applications, leaving users vulnerable to Man-in-the-Middle attacks (MITMAs) (Fahl et al., 2013; Conti et al., 2016). SSL certificates are applied to encrypt data in transit between the host -either the server or the firewall- and the user, through the internet browser. With the use of SSL certificates, it is ensured that the information transmitted is delivered from/to the appropriate server without interceptions. Some types of SSL certificates such as organization SSL or extended validation SSL add an additional layer of credibility since the visitor of a website may see the organization's information knowing that is a genuine entity (Bhiogade, 2002).

From a total of 73 libraries that have a website, as shown in Table 4, we have identified that only 25 use SSL in order to secure the transmitted data.

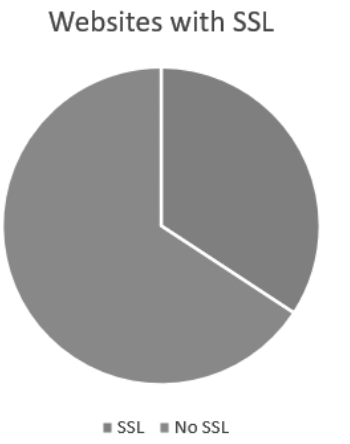
Despite that, the NLG Network was created with the aim to assist academic, research, public, civil, and school libraries to develop and evolve —among others— their online services provided to all interested parties, since the usability and security of these services are weak—if not embryonic—at least.

In order for the NLG Network to participate in national and/or international financial programs, it must previously invest in both security and data protection, especially after the application of the GDPR. Following the requirements of international laws and standards, the GDPR has come to set new rules regarding how to manage personal data on websites, giving explicit options to each user whether to allow or not the use of their data. An investment in website-security and data protection for the NLG Network and each library member of this network should consider the deployment of

elements of a GDPR-compliant website, the minimum number of which is referred below hereto.

Table 4. The Greek Libraries Network of the National Library of Greece includes 73 libraries that have a website. From a total of 73, 17 libraries have updated Privacy Policies and 25 use SSL in order to secure the transmitted data.

	Libraries
Privacy Policy	17
No Privacy Policy	56



	Libraries
SSL	25
No SSL	48

4.5 Elements of GDPR-compliant websites

An interesting factor regarding the GDPR is the enforcement of fines to the organizations that appear to be non-compliant. The GDPR Enforcement Tracker[100] is a website that contains a list and overview of fines and penalties which data protection authorities within the EU have imposed under the GDPR the past couple of years. No library appears to have been penalized, so far. Since some fines are not made public, this list is not exhaustive, and thus the GDPR Enforcement Tracker cannot guarantee that libraries Europe-wide have escaped fines for non-compliance with the GDPR.

[100] See GDPR Enforcement Tracker available at URL: https://www.enforcementtracker.com/

Drawing on the available data and by searching for keywords '*web*' and '*email*', we have found 27 cases relate to non-compliance regarding websites and eight cases about marketing email campaigns, until June 2020. While this may not be a large number of penalized entities, combined with the current risk incurred by the growth of personal data processing, more fines are expected to emanate in the future, if the requirements of the GDPR are not fully respected.

In order to consider how the GDPR impacts a website, it should be determined how the website under consideration interacts with organizational activities in the online world such as email campaigns, digital marketing, social media strategies, and further actions that might include personal data. The GDPR makes clear that each user should firstly provide consent regarding the use of their personal data, thus more transparency should be applied regarding the handling of personal data. This consent must be cumulatively (1) freely given, (2) specific, (3) informed, and (4) unambiguous.[101] All these four traits of consent must be explicitly clarified in the content of a Terms and Conditions of Use text on the website of the NLG Network as well as on the individual websites of its library members. The Data Subject Consent Form and the Data Subject Consent Withdrawal could be linked documents/forms to the Terms and Conditions of Use text. In the event of underage users of the library website and/or applications used through it, the Parental Consent Form and the Parental Consent Withdrawal Form could be also linked to the Terms and Conditions of Use text. An Access Control Policy may also be linked to the Terms and Conditions of Use text informing users of the library and/or users of the NLG Network about the basic principles for accessing all available information systems, networks, and services and the user-registration process for accessing said means. In the event of permission to use the user's own devices within the premises of a library, the Terms and Conditions of Use text could also include a link to the Bring Your Own Device Policy that applies in the library.

Additionally, and in consideration of the GDPR, the websites of the library members of the NLG Network should contain an analytic and unique Privacy Policy section, analyzing the types of information that are stored, the way they are collected and

[101] See Art.4(11) of the GDPR.

the purposes of the collection. Also, the Data Retention Policy could be linked to the Privacy Policy, as well as the Cross Border Personal Data Transfer Procedure when applicable. The website of the NLG Network and/or the websites of the library members of said Network may elaborate upon Standard Contractual Clauses for the Transfer to Controllers of personal data and/or upon Standard Contractual Clauses for the Transfer to Processors of personal data in the event they make use of them —Controllers and/or Processors— for the processing of personal data through their operation. The Data Subject Access Request Form could be linked to Privacy Policy, too.

An analytic and unique Privacy Policy section is a sine-qua-non element of a website that is compliant with the GDPR requirements (Degeling et al., 2019). In addition to elements linked to the Privacy Policy of a website as reported above hereto, the following elements could be contained in the Privacy Policy: 1) basic information regarding the library and the types of information that are stored and the way they are collected; 2) apart from the GDPR, there should be a reference to applicable national laws like data protection laws and regulations that remain in effect in Greece; 3) if any third-party providers are leveraged upon, they should be mentioned in the Privacy Policy including all services that might track end-user data e.g. Google Analytics, Facebook Analytics, even plugins and applications that use or store such data, etc.; 4) the Privacy Policy should include detailed information regarding the personal data gathered by the website and/or the applications offered to users through it as well as the purpose for each personal data process; 5) a detailed action plan in an event of a data breach or a successful hacking attack; for example, if any data is in danger, the data subject should be notified following a concrete communication strategy of which the core elements may be reported in the Privacy Policy; 6) the Privacy Policy should include detailed information about the data controller and the data protection officer, including contact details of both of them (Lindèn et al., 2020; Zuiderveen Borgesius et al., 2017).

Furthermore, all websites should contain a Cookies Policy which includes a detailed list of the cookies that are collected by each website including the data of the visitors. Following the principles of the GDPR, both the NLG Network’s website, as well as each library member's website, should contain a detailed list of the cookies that are collected by each website including the data of their visitors. Also, there should be

included a notification banner or page regarding the cookies, in order for the user to choose the cookies that they want to provide in any case (Sanchez-Rola et al., 2019; Dabrowski, 2019).

A Privacy Notice composed in lay terms should be provided on the index page of websites of both the NLG Network as well as on the individual websites of its library members.

Regarding plugins and further applications that may be implemented on the website of any library member and/or the website of the NLG Network, they should be compliant with the GDPR. If the application or plugin is created by a third-party, then the website owner (library or the NLG Network) should check whether the third party is GDPR-compliant or not. In any case that any plugin or application is not compliant with the requirements of the GDPR, then an alternative plugin or application should be found and used.

All checkout pages should follow the GDPR's principles. In order to do so, checkboxes should be included for the consent of the user and link directly to the website's Privacy Policy during checkout.

The Privacy Policy and the Terms and Conditions of Use text, at least, should include among other information specific reference to the user's rights furnished through the GDPR and also include a link to the Data Subject's Rights Form through which a user may apply their rights.

Regarding email marketing campaigns and newsletters, the website of any library member and/or the NLG Network, should include an option for the user to unsubscribe from the mailing list and another option to opt-in the user and check if they gave the consent to store the personal data. Moreover, after a series of marketing email campaigns are sent, the administrator of the website should proceed with the deletion of the low bounces email accounts. In any case of a user request regarding personal data, a reply should be sent within two days. In addition, in any case of a user request regarding deletion or update of data, the library member or the NLG Network itself should reply and act accordingly within 30 days after the submission of the user's request.

There are other documents and procedures which should not be necessarily linked to the texts described in the websites of the NLG Networks as well as in the websites of its corresponding library members, such as an Appendix Inventory of the Processing Activities, the Employee Privacy Policy, the Data Breach Response and Notification Procedure, the Anonymization and Pseudonymization Policy, the Policy on the Use of Encryption, the Information Classification Policy—if any, necessary—, the Mobile Device and Teleworking Policy, the Clear Desk and Clear Screen Policy, the Supplier Data Processing Agreement, the Employee Data Retention Policy, the Security Procedures for the IT Department, the IT Security Policy, the Data Breach Notification Form to the Supervisory Authority, the Data Breach Notification Form to Data Subject.

In the event of a hacking attempt, website backups are crucial. Although they should not be regarded as a substitute for having in place website security infrastructure, a backup can always help restore damaged files, reassuring the fast recovery of the library's image to the online world. All library members as well as the NLG Network should have a backup solution for their websites in order to be compliant with the GDPR and also in order to ensure functionality and availability of their resources. Nevertheless, the administrator of each website should ensure that there are not more than three backups that may include the personal data of users. A proper backup solution should be, firstly, off-site. In case the backup is stored in the same server where the website is hosted, it is as vulnerable to attacks as the website itself, and this is a major security risk. The backups should be stored off-site in order to be immune to hacking attempts or a potential hardware failure. In addition, backups should be automated, taking advantage of the numerous backup solutions available. Backups should be checked periodically in order to ensure that they function in a proper way. Furthermore, the existing backup should be encrypted with a strong encryption algorithm and only the administrator of the website should have the ability to download the backup. It is important to possess a local backup of the whole entity and an external backup not linked to the application in case of an equipment failure or of a malicious episode (Politou et al., 2018).

Nowadays, more and more websites suffer from successful hacking attempts due to outdated or unpatched applications or Content Management Systems (CMS). It is of crucial importance to update each website whenever a new version is released, either

regarding the CMS or regarding the plugins that are installed. As we mentioned above, the majority of hacking attempts targeting websites are automated. Bots scan every site they can to discover any exploitation opportunities. It is not efficient to upgrade once a month or even once a week anymore, because bots are extremely likely to discover a vulnerability before it is patched.

A CMS is a software dedicated to creating and managing content for a website, on a particular platform. Nowadays, it is strongly recommended to use CMS such as WordPress (Lindèn, 2019) and Magento for security, compliance, and functionality reasons. CMS platforms have teams that implement and release functionality and security updates periodically (Patel et al., 2011). A CMS, like WordPress, has a strong community where anyone can share questions and find answers regarding how to further secure a website. While a CMS often provides frequent security updates, the use of third-party extensible components, such as themes or third-party plugins and applications, leads to vulnerabilities that cyber threats can easily target and exploit. The most common hacking attempts targeting websites are completely automated and most of these attacks rely on the default settings of a CMS. Thus, the majority of the attacks can be restrained by configuring the default settings by the website administrator. Nevertheless, despite the fact that a CMS may appear to be more vulnerable because it is based on an open-source framework, a website based on a CMS can reach a satisfying level of security not only regarding compliance matters but also regarding real hacking threats.

One of the most adequate techniques to enhance the security of a website is the use of a web application firewall (WAF). Even though a WAF may assist in order to fulfill the Payment Card Industry Data Security Standards (PCI DSS), it can hardly provide protection with regard to every security incident that may occur online in the case of a library website. There are also other aspects that may affect the security of a website such as the errors that may occur due to mistakes based on the human factor. Furthermore, the use of SSL certificates is not enough in order to prevent a malicious user to get unauthorized access to a library's website. A vulnerability that may exist on a library's website can enable a potential attacker to capture communications and receive personal or restricted data. Moreover, even if a library's website under consideration is completely patched with all the updates, a malicious user may focus the hacking attempt on

distributed denial of service attacks (DDoS) with the aim to slow the websites' response to the legitimate users or even set it not available at all. Thus, a web application firewall is designed in order to prevent a library website from such malicious activities.

4.6 Conclusions

As we have mentioned in past work (Vavousis et al., 2020), the requirements of the GDPR affect not only the websites of the library members of the NLG Network but their day to day work as a whole. In parallel with the creation or re-built of a compliant and secure website, each member of the NLG Network will have to determine the level of compliance for its website and IT infrastructure regarding the processing of personal data. They should also comply with privacy by design and privacy by default principles by deploying technical and organizational means on how to process personal data and adopt privacy-enhancing techniques and policies capable of protecting personal data from malicious activities.

The security of an IT infrastructure of a library member of the NLG Network is an ongoing procedure that requires constant research and changes according to the needs of the library based on international standards, regulations, and new-found solutions. In a similar vein, a library's website security is a never-ending process that needs continuous assessment in order to achieve compliance with applicable laws and regulations such as the GDPR and in order to reduce the risks that may occur, protecting it from internal and external cyber threats. Thus, a library's website protection is certainly a procedure that evolves constantly and an important component for the management of a site (Johnson et al., 2020).

Website security is essential for every organization, especially for the NLG Network and for each library member's online presence. The absence of proper website security can lead to multiple unsolicited events such as traffic losses, website reputation loss, and users' data leaks. The calamities that may occur due to improper website security such as users' data leaks, might lead to litigation enacted from negatively

affected data subjects, the imposition of fines by the Supervisory Authority in Greece for violation of the GDPR and of relevant legislation, and to a tarnished library reputation.

In order to conduct an optimum solution with regard to security and compliance for the websites of the NLG Network and for each library member's online presence, protective mechanisms that ensure high availability and security should be in place, not only for the visitors but also for the IT infrastructure.

The Greek Libraries Network could expand beyond the Hellenic territory and include libraries abroad that serve as a center of Greek culture.

Following the project to provide an ILS to every library in Greece, a common identity management system (IDM) for users of all member libraries (e.g. one ID card for all to unified services) is under consideration. Moreover, a document exchange system is under development to organize and automate collection exchange among Greek libraries. In this context, the provision of legal services and consultancy to all member libraries (e.g. intellectual property rights, collection donations, GDPR, etc.) is also one of the core future developments of the Network.

Chapter 5: A worldwide password analysis for public Wi-Fi infrastructures

5.1 Overview

As the world has become ever more mobile, the need to stay connected, either to maintain personal relationships or do business, is growing for a large number of people. While, in the past, this was done mainly on our personal computers, networking today with the internet of things, has become more portable and accessible than ever before through mobile and smart devices that we carry with us all day and use on public Wi-Fi networks. Global public Wi-Fi hotspot numbers grew from 1.3 million in 2011 to 5.8 million in 2015, marking a 350% increase in just four years. According to a study conducted in a sample of 2,003 USA citizens, the average user checks their mobile phone 47 times per day (Westcott et al., 2019). This number increases between the ages of 18 and 24, where they check their smartphones 86 times per day. Indeed, we rely on our smart devices instinctively, hungry for instant, relevant information. In the age of ubiquitous computing, we often seek convenience and speed.

The craving to stay connected indicates that this will be done from everywhere, blending and overlapping public and private spaces. Free of charge web browsing through public Wi-Fi networks at coffee shops, airports, restaurants, train stations, and schools offers customers the opportunity to connect quickly and in the convenience of their own apparatus (Balasubramanian et al., 2009). However, what happens when users enter the public domain? In other words, what is out there for users? While more and more governments promise widening the availability of public Wi-Fi and public Wi-Fi hotspots are everywhere in big cities, awareness of security issues does not seem to be on a parallel path.

As staying connected has become a daily routine for most people, the number of ensuing security threats is not a matter of concern neither for the users themselves, who seek speed, nor for the administrators who want to avoid lengthy protection procedures.

Nowadays, a malicious user does not need to be adept at social engineering to get unauthorized access to a network or obtain by illegitimate means a Wi-Fi password (Krombholz et al., 2015). While the literature abounds with references to attacks occurring in private Wi-Fi infrastructures (WarDriving, Dictionary attacks, Password Theft and WEP/WPA or WPS attacks), the dangers in public networks are somewhat different as the password is already known. Indeed, uncontrolled access to public Wi-Fi hotspots and robust mobile security often conflict with one another. Users not only run the risk of malware, Trojan or ransomware infections, but any password or login credentials they enter are transmitted in cleartext due to the lack of encryption, making them ideal targets for cybercriminals. Businesses also face the proliferation of legitimate-looking twin networks set up by cybercriminals to steal information.

Despite how easy it has become for an attacker to achieve unauthorized access to a network or a service (Chou et al., 2013; Kelley et al., 2012), security is more than often compromised over the exigency to create passwords that are easy to remember in order to facilitate the usability and speed of the service. While in private spaces the onus is on the individual to provide a secure environment for their data, the security of the available public Wi-Fi networks is a different matter as it affects a large number of mobile users.

In this chapter, drawing on a uniquely compiled vast database of the available public (only) Wi-Fi passwords, we first examine the security level of publicly available networks, by analyzing the complexity and strength of the passwords. Public Wi-Fi networks, while certainly useful and increasingly available, have numerous pitfalls. On their own, they pose significant security and privacy concerns for users, but in combination with a lack of understanding about the risks, the threats are markedly amplified. Secondly, we compare our data with private password databases from previous research to identify similarities and differences. Based on this analysis we evaluate the solutions offered in the literature and make some recommendations to raise security awareness. Overall, the contributions of this chapter are twofold:

- We provide and analyze for the first time a password database of more than one million passwords of available public Wi-Fi routers from the biggest cities and capitals around the world, employing an innovative methodology. In doing so, we provide new

insights on how businesses that offer public Wi-Fi networks perceive security and ultimately deal with it.

- By comparing our database with other private password datasets, we aim for the first time in the literature to lay down a more rounded idea of how convenience outweighs consequence, especially with how people use their mobile devices and explain security naivety.

The points above seek to show that this research is both necessary and timely. First, it is necessary because the literature has tended to focus mainly on analyzing private password databases. Second, this chapter is timely because, in the age of ubiquitous computing, when people seek to stay connected at all costs, we need to discuss alternatives for the enhancement of the security of public Wi-Fi networks. This is an important time to be conducting new research that focuses on massively used networks across the world.

This chapter will unfold as follows: the following section analyses the number of security threats that come with free and public Wi-Fi and provides an overview of related work. In section 1.3 we describe the methodology employed concerning data collection and analysis. Section 1.4 presents the findings. Section 1.5 explores the impact and section 1.6 identifies key recommendations. The concluding section (1.7) summarizes the above and highlights the importance of further research in this field.

5.2 Background and related work

This section provides a brief overview of the hidden risks that logging onto Wi-Fi networks carries, covering their most critical aspects. Section 1.2.2 outlines relevant research focusing mainly on work with private password databases. A comprehensive analysis of all the related literature requires an extensive review, which is outside the remit of this research.

5.2.1 Overview of security threats

As mentioned above, security threats on public Wi-Fi networks differ considerably from those on private networks as the password to the network is already known, hence access to the network is greatly facilitated. This way, a malicious user could spread any kind of malware, such as Trojans, ransomware, worms, and infect the users within the network. What is more, when users connect to a public Wi-Fi network transmission within the network it is not encrypted; thus, 'authorizing' potential eavesdroppers to capture and access everything that is being sent over the network.

The most common type of Wi-Fi security threats is known as Man-in-the-Middle Attacks. Mobile apps and devices need to connect with remote servers in order to function, often failing to use standard authentication methods; thus, making users susceptible to these attacks. As the name implies, a malicious actor can go between a user and a network and gain access to and exploit private information, sent to a bank or an online store, for example. The intruder can also identify a person's location, gain access to personal messages, and access stored information within the device. Similarly, sidejacking allows attackers to intercept or log data flowing around public hotspots, to steal a session cookie containing usernames and passwords from a variety of websites, such as Facebook or LinkedIn.

These attacks are far from being the only ones though. Evil ad hoc networks, known as evil twins, can turn up wherever there are public Wi-Fi hotspots and can be used to trick unsuspecting users into connecting to them. Evil twins are rogue Wi-Fi access point that look legitimate or perhaps even mimic a trusted network. By connecting to them, users unintentionally expose passwords and sensitive information, like credit cards, to attackers. Public Wi-Fi networks are extremely vulnerable to this type of attack. While the above attacks may appear more complicated, there are others that initially seem harmless, such as packet sniffing. Packet sniffers can be used to capture traffic on an entire network or parts of it, which can then be analyzed to detect network-related problems, which can also assist network monitoring and management. However, this can also be abused by attackers to capture all the data sent over a public wireless connection, including login details, passwords, and user's cookies.

The above shows that uncontrolled access to public Wi-Fi hotspots and robust mobile security often conflict with one another. Indeed, it is not a matter of whether there is going to be an attack on publicly available networks, but when, how, and at what costs to users. As our need for constant connectivity will not diminish over the years, but instead is likely to increase out of proportions, the use of public Wi-Fi networks will also stay on the rise. Therefore, the need to build security walls around naïve users who are on the move is today more urgent than ever. However, before we do that, we need to understand how public networks operate, their vulnerabilities, and their limits. In the following section, we will briefly discuss relevant academic work to explore how Wi-Fi security has been approached in the literature.

5.2.2 Related work

To the best of our knowledge, there is no relevant work that deals specifically with the analysis of passwords that are used in public Wi-Fi hot spots. In the age of ubiquitous computing, there has been understandably a wealth of research that looks at how people connect, from where and with what costs. Most of this literature focuses on analyzing how users act in their private spaces and the dangers they encounter regarding Wi-Fi security, drawing mainly on the analysis of leaked password databases. For the purposes of this chapter, we will refer to the work being conducted in this area. Hence, this section will briefly present previous research on private password choices and Wi-Fi security, while at the same time carve the context of our work, which offers for the first time a unique database of available public (only) Wi-Fi passwords, evaluates their strength and aims to lead the way for future research in the field.

In their work, P.G. Kelley et al. (2012) measured password strength by simulating password-cracking algorithms to examine the effects of composition policies on the guessability of passwords. They developed an efficient distributed method for calculating the effectiveness of several heuristic password-guessing algorithms. This method works as a guess-number calculator by which they can check whether or not a password-guessing algorithm trained with a given data set would perform this function. In doing so, they compared the guessability of passwords under different password-composition

policies. According to their findings, a password-composition policy which requires a long sequence of characters without any additional restrictions provides effective resistance to guessing. For example, they claim that a basic16 password methodology is more effective than comprehensive8, despite the US National Institute for Standards and Technology (NIST) considering them as equally competent (Komanduri et al., 2011).

J. Bonneau (2012) analyzed the Yahoo! leak, a corpus consisted of approximately 70 million passwords. In order to examine such an enormous corpus, Bonneau did not use metrics such as Shannon entropy ("Shannon entropy", 2020) and guessing entropy ("Guessing entropy", 2020); instead, he developed partial guessing metrics including new variants of guesswork parameterized by an attacker's desired success rate. More specifically, he was able to formalize improved metrics for evaluating the guessing difficulty of a given distribution of passwords, introducing α-guesswork as a tunable metric that can effectively model different types of practical attacks. He was also able to adopt security metrics that rely only on the statistical distribution of passwords. While this approach requires large data sets, it was effective in eliminating bias from password-cracking software by always modeling a best-case attacker, which allowed the researcher to assess and compare the inherent security of a given distribution. In this respect, Bonneau found surprisingly little variation in guessing difficulty regarding password composition. He concluded that a tighter password policy might produce distributions with significantly higher resistance to guessing.

C. Shen et al. (2016) analyzed the password attributes drawing on a database of approximately 6 million passwords, taking into consideration their length, composition, symbols statistics, letter-case, and selection. Their main goal was to compare research on the analysis of leaked password databases over several years to show how password composition policies have been changed over time. They claim that nowadays apart from the irrefutable absence of public security awareness, there is also a lack of adoption of security enhancement methods in real life. Their general results include some interesting findings, such as the fact that the average password length is at least 12% longer than previous research results and 75% of the corpus under consideration have length between 8 to 10 characters. They also found an increase in using exclusively numbers as passwords. Moreover, they claim that easy to reach symbols such as ".", "@", "!", are on

top of users' choice. Last but not least, they concluded that still many users prefer top common passwords or even names as passwords to access their personal accounts.

Regarding Wi-Fi security, there are numerous works available that present the impact of the absence of Wi-Fi security in general. The most relevant works regarding the current research are analyzed below.

N. Cheng et al. (2013) explored serious issues regarding personal data, including the leakage of private information/data, that travelers face while using public Wi-Fi networks. They collected data from airport databases, which shows that privacy leakage in terms of identity, location, financial, social, and personal privacy, may affect approximately up to 70% of users on the move, highlighting that users are not fully aware of the dangers in entering public wireless environments.

P. Bajpai et al. (2014) employed war dialing to determine the level of security of Wi-Fi routers in two cities in India, Delhi and Indore. War dialing is a technique that automatically scans a list of telephone numbers, usually dialing every number in a local area code to search for modems, computers, bulletin board systems (computer servers), and fax machines ("Wardialing", 2020). Drawing on this technique Bajpal et al. (2014) collected data regarding the number of users who continue to configure weak Wi-Fi security and discussed related security implications. They proposed a simplified solution for every user to help them determine if the Wi-Fi network in use meets standard security requirements. In so doing, they created a program, named 'checkmywifi'. It is built for Windows machines and automatically detects the wireless network adapter on the user's PC, and subsequently identifies the security protocols used among other features. If all the checks are satisfied, it notifies users whether or not they adopt optimum Wi-Fi security protocols according to current standards.

C. Swanson et al. (2010) explored Wi-Fi security and privacy by analyzing users' choices regarding passwords and reluctance to change. Their work mainly focuses on users' choices based on their awareness of the existing vulnerabilities when logged on wireless networks. Relying on a demonstration of system vulnerabilities to 11 users of a Wi-Fi hotspot and interviews with them, they argue that users make security choices based on an analogy to the physical world, failing to develop a realistic view of security

risks. Therefore, they argue that the key to creating better security tool designs is to understand how and why users rationalize actions. Future researchers will, thus, be able to educate users and present information in new, more effective ways.

A. Cassola et al. (2015) focus on the issue of Wi-Fi hotspots in untrusted access points. Their paper proposes a system protocol that allows an internet service provider (ISP) to authenticate its clients. The recommended solution manages to hide the clients' identity from both access points and the service provider at the time of authentication. More specifically, the client is guaranteed that either the provider cannot do better than to guess their identity randomly or they obtain proof that the provider is trying to reveal their identity by using different keys. Their protocol is based on Private Information Retrieval (PIR) with an augmented cheating detection mechanism based on the authors' extensions to the NTRU encryption scheme. PIR is a protocol that allows a user to retrieve an item from a server in possession of a database without revealing which item is retrieved ("Private information retrieval, 2020). The NTRU encryption scheme is a lattice-based alternative to RSA and ECC ("NTRUEncrypt", 2020). Their suggested encryption scheme makes possible the auditing of multiple rows in a single query and optimizes PIR for highly parallel GPU computations with the use of the Fast Fourier Transform (FFT). In their work, they also propose an implementation compatible with the Wi-Fi Extensible Authentication Protocol (EAP) along with optimizations for over 10 million clients. They evaluate the performance of mobile and provider components and show that a client can be authenticated in 43.9 milliseconds on a GPU platform, giving an end-to-end authentication of 1.12 seconds.

In a similar vein, L. K. Raju and R. Nair (2015) analyze the numerous security issues that users face nowadays, due to Wi-Fi access points that have not been well configured so as to ensure the security of the data they manage. The existing solutions to this problem are mostly user-centric and they are viewed as precautions taken from the user side. They propose a security protocol that ensures secure internet access through public Wi-Fi hotspot, providing individual confidentiality during the communication. The solution tries to eliminate the dependency on any pre-shared information between the AP (Access Point) and the client device to implement security. Existing WPA2-PSK protocol is modified to generate an Instantaneous Session Key (ISK) between the client and the

Access Point through secured Diffie Hellman key exchange thereby eliminating the dependency on a pre-shared key.

The aforementioned is of high value because it highlights the numerous problems regarding Wi-Fi security and further shows how users understand and navigate security in their private space. However, to date there has been no academic study of the available public databases, rendering the attitude of users in public unexplored. Drawing on the existing literature, this chapter analyzes a uniquely compiled database of available public (only) Wi-Fi passwords compared with the available private password corpora and examines for the first time whether public Wi-Fi providers create a secure space for users. In doing so, we delineate the world that unwary users in the public enter every day. We hope that this will be the foundation for future research in this field.

5.3 Password Collection and Analysis methodology

To collect the set of public Wi-Fi passwords, we used an Android device where we installed the following mobile applications from Google Play: Wi-Fi Map (2016) and Packet Capture (2016). Wi-Fi Map provides a crowdsourced list of public Wi-Fi passwords (such as restaurants, hotels, cafeterias, etc.) for both Android and iOS devices. Specifically, based on the GPS coordinates of the mobile device, the application shows the passwords of nearby public Wi-Fi networks. The paid version of the application allows for downloading and storing locally, on the mobile device, the passwords available for a given city for later offline access. There are several other similar available mobile applications that provide the same services as Wi-Fi Map, like Swift Wi-Fi: Global Wi-Fi Sharing (2016), Free WiFi Internet Finder (2016), Free WiFi - Wiman (2016), Show WiFi Password (2016), WADA Wi-Fi (2016), Maps - Free Wi-Fi (2016), WiFi Chua Free WiFi password (2016), Wi-Fi Space (2016), and more. The reason that we selected Wi-Fi Map over the aforementioned applications was that it allowed us to collect a large corpus of passwords through its API.

As we mentioned above, to collect the passwords, we employed the paid version of the Wi-Fi Map and another mobile application named Packet Capture. The Packet Capture application was used to capture the HTTP traffic of the Wi-Fi Map app. The captured data were in the form of JSON files, with every JSON file representing a city, including information about the location of the public Wi-Fi network, its SSID name, password, owner id, and more. We then extracted only the information relevant to our analysis. We aimed to study the results of the capital cities from Europe, Asia, Africa, Australia, North and South America, as well as cities that had over 600 registered Wi-Fi networks in the Wi-Fi Map application. This way, we managed to collect 1,176,135 passwords from 191 cities of 111 countries all over the world. Table 1 shows the passwords collected sorted by continent, mentioning the number of cities included from each continent, the number of passwords collected, and how many of those were unique.

Table 9: Total of Passwords. The collected passwords were then analyzed using Pipal (DigiNinja, 2012) and custom scripts. Pipal is an open-source password analyzer implemented in Ruby, that provides statistical results in a fast and accurate manner.

	Africa	**Asia**	**Europe**	**North America**	**South America**	**Australia**	**Total**
Cities	12	63	50	52	12	2	191
Passwords Obtained	98,567	749,326	172,798	92,242	57,571	5,631	1,176,135
Unique Passwords	67,010	464,648	120,738	71,777	45,118	4,548	773,839

5.4 Findings

Apart from providing statistical results for the collected Wi-Fi passwords, we wanted to study whether users tend to employ the same techniques in creating their passwords, even though they are destined for different purposes. To this end, we compare

the results of our study with results that have been derived from the analysis of leaked online account passwords. Recall that online account passwords are used to login to online services and have been leaked in public by hacking incidents. For this purpose, we retrieved statistics from the following incidents:

- The “Chinese Software Developer Network” (CSDN) website hack in late 2011. CSDN is the biggest network for Chinese software developers and, back in December 2011, was hacked via SQL injections, disclosing more than 6 million users’ registration information in plaintext. In 2016, C. Shen et al published the dataset’s analysis results in their paper.

- The “LinkedIn” data breach in 2012. LinkedIn, the most prominent business-oriented social network, suffered a major data breach, exposing 177 million emails and unsalted SHA1 passwords (“LinkedIn Revisited - Full 2012 Hash Dump Analysis”, 2016). Four years later, in 2016, the leaked database was made available online and many security researchers were working on cracking the hashed passwords. Today, 100% of the passwords have been cracked by the KoreLogic Security team, among others, that also provided us with their final results.

- The “Yahoo! Voices” online publishing service, a division of the international Yahoo! Web service provider, hacked in 2012 via an SQL injection attack (Lunden, 2012). 442,832 passwords stored in plaintext were disclosed and today can be found online.

The section 1.4, where all findings are presented and discussed, is organized as follows. First, the password policies and limitations are discussed in sub-section 1.4.1. In sub-section 1.4.2 we list the Top 50 used passwords resulted from our study. Sub-sections 1.4.3 and 1.4.4 show, respectively, our findings regarding password length and the Top 10 base words (words that cannon be broken down into smaller units) that are most commonly employed by users in their passwords. Part 1.4.5 presents the statistics for the last digit, while, in parts 1.4.6 and 1.4.7, the usage of character sets and symbols are examined. Lastly, sub-section 1.4.8 includes further findings. In all the aforementioned sections, the results produced by our work are presented next to those of CSDN, LinkedIn, and Yahoo! Voices, focusing on the similarities, differences and trends exhibited, ultimately trying to answer the question of whether users carry their password

composition habits observed in account passwords, while also choosing their Wi-Fi passwords.

5.4.1 Password Policies and Limitations

To better understand the conditions under which the passwords were chosen, we need to take into consideration the password policies or the lack of them. Users tend to choose simple passwords, easy-to-remember, therefore, easy-to-guess. Oblivious to the minimum complexity requirements that would qualify a password as "strong", it comes as no surprise that users would rather trade their account's security for their personal convenience. While it may not be the remedy to the problem, service providers are in a position of enforcing complexity requirements by employing well-established password policies. It's worth noting again the diversity of our data sources since we compare passwords from three major web platforms (CSDN, LinkedIn, Yahoo!) against public Wi-Fi passwords set by, most likely, individual, non-expert users. Password policies would make sense in the context of the former, but there is no use in setting strong composition rules for a password that is not meant to be kept secret, as it is the case for public Wi-Fi network passwords. The only limitation that can be applied to a Wi-Fi password is regarding its length and it's based on the security protocol used by the network. If the network employs the Wired Equivalent Privacy (WEP) protocol, the password must be at least 5 characters long, while for Wi-Fi Protected Access (WPA) and Wi-Fi Protected Access II (WPA2), the minimum length is 8 characters.

The rationale behind LinkedIn's password policy, back in 2012, did not differ from that of wireless security protocols described above. The only restriction set by the social network was for the chosen password to be at least 6 characters long, with no complexity requirements imposed, as the plainness of the most common LinkedIn passwords suggests (Table 2). Today, LinkedIn does not accept "password" as a user's password, but will still allow the use of "p@ssword". CSDN and Yahoo! Voices, on the other hand, at the time of the attacks did not enforce any kind of restrictions over their users´ choices of password, as it can be proved by the fact that, in both leaked databases, single-character passwords were found. However, with the increasing rate of data breaches

globally, no password policy, no matter how strict it is, can protect the users after an incident, if the provider stores their credentials in plaintext, like in the cases of CSDN and Yahoo! Voices.

5.4.2 Top Passwords

Table 2 lists the most common passwords with their respective frequency rate in each dataset, in the form of percentages, as those were derived from the analysis process. The first column presents the results of our work, showing the Top 50 passwords used for accessing public Wi-Fi networks worldwide. The list predominantly consists of passwords that are exclusively composed of numeric sequences in ascending and/or descending order, as well as repetitive patterns of a single or more digits. The first non-numeric password makes its appearance at the 23rd place on the list and it's the word "password", which is also one of the passwords all Top 50s' have in common, along with "123456", "12345678" and "123456789". These results are far from unexpected considering the singular nature of passwords created for public Wi-Fi infrastructures. Shared passwords are not meant to provide security. They only serve usability purposes, which perfectly explains their rather simple structure.

The same cannot be said, however, for the CSDN passwords, listed next to ours in Table 2. According to the work of C. Shen et al. (2016), the Chinese network demanded from every registered user to submit their personal information (gender, profession, education, and working experience), and therefore we would normally expect to see stronger passwords on the CSDN Top 50 list. Interestingly, the CSDN Top 50 list shares almost half its entries (23/50) with the respective list of Wi-Fi passwords, exhibiting a similarity rate far greater than the one exhibited against the LinkedIn and Yahoo! Voices lists, which, on the other hand, appear to have 50% of their entries in common.

Another observation worth discussing is the trend of keyboard patterns (Schweitzer et al., 2009) used as passwords. Users, in their effort to create complex passwords, often resort to employing patterns of keys on the keyboard that, while seemingly creating random strings, are easy to remember. Attackers, however, are well aware of this

password-creating technique and have already embodied character combinations derived from keyboard patterns into their dictionaries used for dictionary attacks (Chou et al., 2012), rendering the keyboard-based passwords insecure. All Top 50 lists above (Table 2) include “qwerty”, which is probably the most common keyboard pattern, either in this form or in variations (“qwertyui”, “qwertyuiop”). At least 11 passwords on the CSDN Top 50 list can be characterized as keyboard-based. Three of them (“147258369”, “789456123”, “123654789”) do not make sense in the context of the classic qwerty keyboard, but they create a movement pattern when typed using the numeric keypad (numpad) layout.

Table 10: Top Passwords Comparison (continues on page 90).

	Comparative Analysis							
	Current	**%**	**CSDN**	**%**	**LinkedIn**	**%**	**Yahoo! Voices**	**%**
1	12345678	2.65%	123456789	3.66%	123456	1.81%	123456	0.38%
2	123456789	1.64%	12345678	3.31%	Linkedin	0.33%	password	0.18%
3	1234567890	0.87%	11111111	1.19%	password	0.30%	welcome	0.10%
4	12345	0.51%	dearbook	0.72%	123456789	0.24%	ninja	0.08%
5	123456	0.47%	00000000	0.54%	12345678	0.15%	abc123	0.06%
6	1122334455	0.36%	123123123	0.31%	111111	0.14%	123456789	0.05%
7	11223344	0.33%	1234567890	0.28%	1234567	0.12%	12345678	0.05%
8	11111111	0.33%	88888888	0.23%	654321	0.08%	sunshine	0.05%
9	88888888	0.28%	111111111	0.11%	qwerty	0.08%	princess	0.05%
10	00000000	0.24%	147258369	0.09%	sunshine	0.08%	qwerty	0.04%
11	12341234	0.22%	987654321	0.09%	000000	0.08%	writer	0.04%
12	1234567	0.21%	aaaaaaaa	0.08%	abc123	0.07%	monkey	0.04%
13	1234512345	0.20%	1111111111	0.08%	charlie	0.06%	freedom	0.04%
14	87654321	0.18%	66666666	0.08%	666666	0.05%	michael	0.04%
15	123123123	0.16%	a123456789	0.07%	123123	0.05%	111111	0.04%
16	1234554321	0.15%	11223344	0.06%	linked	0.05%	iloveyou	0.03%
17	987654321	0.15%	1qaz2wsx	0.06%	1234567890	0.05%	password1	0.03%
18	0123456789	0.15%	xiazhili	0.06%	maggie	0.05%	shadow	0.03%
19	11112222	0.14%	789456123	0.06%	princess	0.05%	baseball	0.03%

20	0987654321	0.13%	password	0.05%	michael	0.05%	tigger	0.03%
21	20152015	0.13%	87654321	0.05%	iloveyou	0.04%	1a1a1a1b	0.03%
22	12344321	0.12%	qqqqqqqq	0.05%	121212	0.04%	success	0.03%
23	password	0.11%	000000000	0.05%	daniel	0.04%	blackhatworld	0.03%
24	1020304050	0.11%	qwertyuiop	0.05%	222222	0.04%	jordan	0.03%
25	1111111111	0.10%	qq123456	0.05%	welcome	0.04%	whatever	0.02%
26	10203040	0.10%	iloveyou	0.05%	baseball	0.04%	michelle	0.02%
27	55555555	0.09%	31415926	0.05%	password1	0.04%	dragon	0.02%
28	12345678910	0.08%	12344321	0.05%	555555	0.04%	purple	0.02%
29	20002000	0.08%	0000000000	0.04%	buster	0.04%	superman	0.02%
30	100200300	0.08%	asdfghjkl	0.04%	shadow	0.04%	1234567	0.02%
31	20152016	0.08%	1q2w3e4r	0.04%	bailey	0.04%	ashley	0.02%
32	12121212	0.08%	123456abc	0.04%	monkey	0.04%	associated	0.02%
33	0000000000	0.08%	0123456789	0.04%	pakistan	0.04%	123123	0.02%
34	qwertyuiop	0.07%	123654789	0.04%	abcdef	0.04%	babygirl	0.02%
35	77777777	0.06%	12121212	0.04%	summer	0.03%	ginger	0.02%
36	00000	0.06%	qazwsxedc	0.04%	chocolate	0.03%	maggie	0.02%
37	111111111	0.06%	abcd1234	0.04%	hannah	0.03%	computer	0.02%
38	112233445566	0.06%	12341234	0.04%	777777	0.03%	trustno1	0.02%
39	99999999	0.06%	110110110	0.04%	7777777	0.03%	football	0.02%
40	aaaaaaaa	0.06%	asdasdasd	0.04%	george	0.03%	cookie	0.02%
41	19901990	0.06%	22222222	0.03%	thomas	0.03%	blessed	0.02%
42	999999999	0.06%	123321123	0.03%	ginger	0.03%	jasmine	0.02%
43	10002000	0.06%	abc123456	0.03%	Linkedin	0.03%	samantha	0.02%
44	admin	0.06%	a12345678	0.03%	freedom	0.03%	pepper	0.02%
45	22222222	0.06%	123456	0.03%	harley	0.03%	charlie	0.02%
46	123123	0.05%	123456123	0.03%	michelle	0.03%	justin	0.02%
47	abcd1234	0.05%	a1234567	0.03%	pepper	0.03%	money	0.02%
48	11111	0.05%	1234qwer	0.03%	letmein	0.03%	writing	0.02%
49	qwerty	0.05%	qwertyui	0.03%	sophie	0.03%	654321	0.02%
50	20142015	0.05%	123456789a	0.03%	zzzzzzzz	0.03%	nicole	0.02%

Table 11: Encryption Methods for Wireless Networks.

	WEP 40 bits	**WEP 104-128 bits**	**WPA**	**WPA2**
Password Length	5 characters	13 characters	8-63 characters	8-63 characters
Key Size	40 bits	104/128 bits	128 bits	128 bits
Cipher	RC4	RC4	TKIP/RC4	AES
Key Management	None	None	EAP Based	EAP Based

Table 12: Password Length Comparison.

Password Length	Current	CSDN	LinkedIn	Yahoo! Voices
1-3 characters	0.42%	0.01%		0.12%
4 characters	0.48%	0.10%		0.63%
5 characters	4.17%	0.51%		1.21%
6 characters	5.68%	1.29%	10.82%	17.99%
7 characters	3.78%	0.26%	13.32%	14.82%
8 characters	**30.68%**	**36.38%**	**29.24%**	**26.9%**
9 characters	**14.06%**	**24.14%**	**16.86%**	**14.89%**
10 characters	**19.24%**	**14.48%**	**13.9%**	**12.36%**
11 characters	7.29%	9.78%	6.54%	4.79%
12 characters	4.87%	5.75%	3.86%	4.9%
13 characters	3.16%	2,61%	1.84%	0.6%
14 characters	1.83%	2.41%	1.01%	0.34%
15 characters	1.14%	1.17%	0.53%	0.19%
16 characters	0.92%	0.77%	0.32%	0.13%
17-34 characters	2.34%	0.32%	-	0.15%
1-34 characters	99.99%	100.00%	98.24%	100.00%

5.4.3 Password Length

In web accounts, enforced password policies dictate the minimum acceptable length for creating a password. In the case of Wi-Fi networks, the minimum password length is set by the employed security protocol.

As stated earlier, LinkedIn obliged its users to create passwords at least 6 characters long.

The analysis of the passwords' length has provided some very interesting findings. To begin with the current analysis, although most passwords (63.98%) were right at the Wi-Fi Protected Access known as WPA ("Wi-Fi protected access", 2020), password limit (i.e. from 8 to 10 characters long), there is still a 14.53% that reflects on Wired Equivalent Privacy known as WEP ("Announcing Our Worst Passwords of 2015", 2016), with length up to 7 characters. It has also been observed that a high percentage of passwords (30,68%) are right on the limit of WPA password strength. This makes it obvious that the administrators of such Wi-Fi networks want to help the user connect to

the available Wi-Fi infrastructure without considering the security parameter. The same applies to the 10 characters' password choice, but with a notable exception. In Table 3, we can observe the encryption methods for wireless networks. We expected that due to the administrators' preference for easier internet access for the user, the research will show 9-digit passwords in second place rather than those with 10 characters. In most countries, however, 10-digit telephone numbers are used (Joshi et al., 2009), so it is very easy for someone to remember a telephone number rather than a random 8 or 9-digit number. Furthermore, the telephone number/password is available in a shop's price catalog, on business cards, and so on and so forth. Moreover, numerous similarities with the results of other researchers regarding password patterns used have been observed. As described in section 1.2.2, C. Shen et al. (2016), who analyzed password attributes on a database of approximately 6 million passwords, have concluded that 75% of the corpus under consideration have a length between 8 to 10 characters, as depicted in Table 4. In a similar vein, our analysis found that more than 63% of our password databases collected have a length of 8 to 10 characters too (Table 4). On the other hand, in the current analysis, it is observed that the passwords up to 7 characters long represent a percentage of 14.53, while in the work of C. Shen et al. (2016) the percentage is only 2.17%.

Comparing the current analysis with CSDN, LinkedIn, and Yahoo! Voices, an interesting fact is that LinkedIn didn't allow passwords with less than 6 characters. Nevertheless, the high percentage of 84.14% covers the passwords composed of 6-10 characters, revealing the users' trend for easy-to-remember passwords. A very interesting observation regarding the Yahoo! Voices database, is that a high percentage of 32.81% is right below the limit of WPA password strength, covering passwords composed of 6 and 7 characters.

As mentioned above regarding the current database analysis, it can be claimed that the administrators' choice is easier access over any security issues. Finally, it is worth noting that, as expected, the prevalence of the quota drops drastically the longer the passwords are, for all four compared datasets.

5.4.4 Top Base Words

More than half of the total password corpora (749.326) were located in Asia and most of them in the area of the Middle East. This explains the prevalence of two common Arabic names “ahmed” and “ahmad” in the top 10 base words analysis for the database collected in the current research. On top of the list, however, again was “password”, followed by sequential patterns here too, such as “abcd”, “qwerty”, “qwertyuiop” etcetera. Compared with previous research on passwords for private online accounts (Adam, 2015), as shown in Table 5, we can identify certain similarities. In both lists, the words “password” and “qwerty” are among the top 3 results with the rest being somewhat different but based on the same premise of base-phrases easy to remember.

It is also observed that passwords from the LinkedIn database also contain names, but with base words such as “alex”, “mike” and “john”. Furthermore, for both current and Yahoo! Voices datasets the password “password” holds the first position.

5.4.5 Last digit

Table 6 presents the most frequently used digit at the end of a password.

In the current analysis, there are no differences found regarding the last digit, compared to the three databases of CSDN, LinkedIn, and Yahoo! Voices. An interesting finding though, is that digit “1” holds the first place for both the LinkedIn and Yahoo! Voices leaks, while in the current database and for CSDN it holds the fourth place.

Table 13: Top Base Words Comparison.

	Current	CSDN	LinkedIn	Yahoo! Voices
1	password	dearbook	linkedin	Password
2	ahmed	wang	link	Welcome
3	abcd	abcd	love	Qwerty
4	qwerty	zhang	ever	Monkey
5	admin	love	linked	Jesus
6	qwertyuiop	woaini	life	Love
7	ahmad	chen	alex	Money
8	aaaaaaaa	aaaaaaaa	mike	Freedom
9	aaaa	yang	pass	Ninja
10	Guest	qwer	john	Writer

Table 14: Last Digit Comparison.

Current		CSDN		LinkedIn* (unique)		Yahoo! Voices	
Digit	**%**	**Digit**	**%**	**Digit**	**%**	**Digit**	**%**
0	9.64%	8	10.64%	1	10.06%	1	10.54%
3	7.76%	9	9.92%	3	6.95%	3	6.60%
5	7.70%	0	9.59%	2	6.80%	2	5.56%
1	7.60%	1	9.48%	0	6.73%	7	4.61%
8	7.41%	3	7.66%	7	5.93%	9	4.50%
4	6.31%	6	6.88%	9	5.87%	6	4.04%
9	6.26%	2	5.69%	5	5.72%	8	4.03%
2	5.18%	5	5.61%	8	5.62%	4	4.00%
6	5.14%	4	5.60%	4	5.61%	0	3.96%
7	4.34%	7	5.16%	6	5.43%	5	3.93%

5.4.6 Character Sets

Drawing on our list, a percentage of 40.54% covers passwords containing numeric characters only, while in the CSDN database the same percentage is 45.01%. While the high prevalence of numeric characters seems to be a particularity of our research, as previous studies have found them to be rather uncommon (4%) (Troy Hunt, 2011), the frequency of alphanumeric characters is common in both studies. In parallel, the same percentage for LinkedIn and Yahoo! Voices leaks is 8.86% and 5.89%, respectively.

Regarding the alphabetic passwords in the current analysis, the percentage is 20.99% while in the CSDN it is 12.35%, being the lowest among all four. The same percentages for LinkedIn and the Yahoo! Voices analysis are 23.17 and 34.67%. It is worth mentioning that the alphabetic passwords are, on their vast majority, lowercase letters, as shown in Table 7. An interesting finding is that regarding alphanumeric passwords the LinkedIn analysis has the highest percentage (60.22%) of such passwords in total. For the same category, the current analysis has 31.62%, the CSDN 39.36%, and the Yahoo! Voices 56.63%. Again, lowercase alphanumeric passwords represent the vast majority of passwords. Following the comparison based on character sets, the passwords including special characters are the fewest in all four databases. Regarding the current analysis, this percentage is 5.64%, for the CSDN it is 3.29%, for the LinkedIn it is 6.25%, which is the highest percentage, and lastly, for the Yahoo! Voices analysis it is 2.34%.

Table 15: Character Sets Comparison.

Character Sets	Current	CSDN	LinkedIn	Yahoo! Voices
Numeric	40.54%	45.01%	8.86%	5.89%
Alphabetic	20.99%	12.35%	23.17%	34.67%
- loweralphabetic	18.45%	12.02%	20.61%	33.11%
- upperalphabetic	0.74%	0.32%	0.86%	0.40%
- mixedalphabetic	1.80%	0.01%	1.70%	1.16%
Alphanumeric	31.62%	39.36%	60.22%	56.63%
- loweralphanumeric	23.83%	35.91%	48.26%	50.61%
- upperalphanumeric	3.81%	2.89%	1.91%	0.77%
- mixedalphanumeric	3.98%	0.56%	10.05%	5.25%
Including Special	5.64%	3.29%	6.25%	2.34%
- specialnumeric	1.08%	0.65%	0.17%	0.04%
- loweralphaspecial	1.09%	0.41%	1.01%	0.40%
- upperalphaspecial	0.12%		0.04%	0.01%
- mixedalphaspecial	0.42%		0.29%	0.09%
- loweralphaspecialnumeric	1.63%	2.23%	2.75%	1.16%
- upperalphaspecialnumeric	0.36%		0.12%	0.04%
- mixedalphaspecialnumeric	0.76%		1.87%	0.60%
- only special	0.18%	0%	-	0%

5.4.7 Symbols Usage

The analysis of the symbol usage of the current research work has provided some very interesting findings in comparison with the work of Shen et al. (2016). To begin with, it has been observed, as it is depicted in Table 8, that in the two top 10 lists of the symbol's usage, 8 out of 10 are the same for all four lists. Furthermore, the two first places are held by the same characters "@" and "." both in our password collection and the CSDN collection, but with reversed order in each list. In a similar vein, the passwords from the LinkedIn password leak and the Yahoo! Voices password leak, the two first places are held by the same characters "@" and "!", but again with reversed order for each list.

Table 16: Symbol Usage Comparison.

	Current		CSDN		LinkedIn* (in unique)		Yahoo! Voices	
	Symbol	%	Symbol	%	Symbol	%	Symbol	%
1	@	31.17%	.	34.57%	@	22.96%	!	26.70%
2	.	16.82%	@	25.43%	!	17.61%	@	21.29%
3	-	15.73%	!	10.92%	.	11.68%	_	12.37%
4	Space	11.89%	*	9.19%	*	8.43%	$	12.12%
5	_	8.83%	_	7.81%	#	7.84%	#	10.50%
6	#	5.52%	#	6.62%	_	7.78%	*	9.79%
7	!	5.18%	+	5.47%	$	7.55%	-	7.11%
8	$	4.30%	/	3.36%	-	6.07%	;	4.69%
9	?	4.20%	$	3.32%	Space	3.81%	&	4.49%
10	*	3.78%	?	2.69%	&	2.77%	.	3.55%
11	:	2.10%	&	2.52%	+	1.97%	%	2.66%
12	/	1.95%	=	2.09%	/	1.93%	+	1.77%
13	&	1.92%	%	2.00%	%	1.87%	^	1.45%
14	)	1.71%	Space	1.46%	?	1.60%	=	1.24%
15	+	1.47%	)	1.41%	,	1.45%	?	0.71%
16	(	1.46%	~	1.34%	)	1.01%	/	0.56%
17	,	1.07%	(	1.29%	=	0.97%	~	0.55%
18	%	0.91%	^	1.23%	(	0.89%	,	0.46%
19	‘	0.77%	;	1.10%	;	0.66%	(	0.40%
20	;	0.69%	-	0.96%	‘	0.62%	)	0.34%
21	"	0.50%	,	0.86%	^	0.59%	[	0.16%
22	=	0.43%	]	0.55%	:	0.49%	]	0.13%
23	^	0.39%	[	0.55%	]	0.28%	|	0.06%

24	<	0.12%	>	0.42%	~	0.27%	\	0.04%
25	|	0.12%	‘	0.42%	"	0.27%	{	0.02%
26	>	0.11%	<	0.36%	[	0.26%	}	0.02%
27	]	0.10%	\	0.32%	<	0.25%	Space	0.00%
28	[	0.08%	:	0.30%	>	0.17%	:	0.00%
29	\	0.07%	{	0.09%	\	0.15%	‘	0.00%
30	~	0.07%	}	0.08%	|	0.05%	"	0.00%
31	{	0.04%	"	0.07%	{	0.05%	<	0.00%
32	}	0.04%	|	0.05%	}	0.05%	>	0.00%

5.4.8 Further findings

Apart from the aforementioned, we would also like to highlight further issues that arose in the overall analysis of all the information gathered. Firstly, only 773,839 from a total of 1,176,135 passwords collected were unique. Secondly, a large number of shared publicly available Wi-Fi networks were identified in Northern Africa, and especially Morocco and Egypt, which are the main touristic attractions in the continent (“WADA – WiFi Maps”, 2016). Given that data roaming is rather expensive in these countries (“Most Expensive Countries for International Roaming”, n.d.), tourists prefer the free of charge available internet, thus explaining the abundance of public Wi-Fi networks. Thirdly, more than half of the available public Wi-Fi networks (749,326) were located in Asia, which could be attributed to the fact that the internet there is more expensive than in Europe and North America (Prince, 2018). Fourthly, it is remarkable that numerous publicly available passwords of approximately 90,000 shared Wi-Fi networks were found in Iraq, where warfare still was taking place at the time of the collection of the current database, possibly reflecting the need for communication with the rest of the world in times of need. Moreover, according to deeper analysis it has been found for the capital of Greece, the city of Athens, 812 passwords of a total corpus of 11662 passwords (6,96%) were combined by Greeklish characters (“Greeklish”, 2020), which is the Greek language written using the Latin alphabet. It has been found, due to the Greek origin of the researchers, that not few public Wi-Fi password administrators choose Greek words written in English.

With regard to the technical side of the analysis, several cases of WEP security protocol use have been observed, despite its numerous security shortcomings (Hulton, 2002), and the existence of other alternatives (e.g. WPA and WPA2 protocols). In the top passwords used, as shown in Table 2, two of them have been located in the fourth and fifth place, which reveals that most probably WEP security protocol is used, based on the length of the passwords ("Wired Equivalent Privacy", 2020). In the cases that less than eight characters are used as a password, routers don't use automatic padding from lab experiments on ZTE models. Therefore, most probably the passwords that are less than 8 characters are used for WEP encryption. Assuming that up to 7 characters are for WEP, we have found 170,892 passwords within that range.

5.5 Impact

The previous section emphasizes that functionality and usability are preferred to security when it comes to offering internet connection to the wider public. This raises a number of questions. For example, who exactly will be able to see the traffic within a given network? Can the physical location of a specific device be tracked? To what extent is it possible for an intruder to take control of the public Wi-Fi infrastructure and manage its traffic? These questions inevitably lead to our main discussion point; that is, how security-aware internet users that use publicly available Wi-Fi infrastructures for their online activities are.

For most of the internet users, it seems harmless to be connected to a publicly available Wi-Fi infrastructure. A user simply requests for the Wi-Fi password and is logged in within a couple of minutes to the local public network. But, if it is considered that a user might be malicious and has access to the public Wi-Fi infrastructure that is also used by the other customers of the same place, the malicious user could become man-in-the-middle, between the communications of the other customers logged in to the same Wi-Fi network. Malicious users, in order to eavesdrop on certain communications, use man-in-the-middle attacks (MitM) ("Man-in-the-middle attack", 2020), by which they masquerade themselves as the authorized receiver of the data/communication. Unfortunately, this is the best-case scenario by which the attacker will have to struggle to

become man-in-the-middle between an end-to-end communication. When there is end-to-end encryption established such as Transport Layer Security (TLS) and Secure Sockets Layer (SSL) (“Transport Layer Security”, 2020), man-in-the-middle attacks are required in order to capture the communication between the two ends. By using TLS/SSL encryption, each end can authenticate one or both parties using a mutually trusted certificate authority, creating a secure environment that an attacker has to be innovative in order to exceed this security measure. To achieve a man-in-the-middle attack in an environment that end-to-end encryption is installed, an attacker should try Secure Sockets Layer Sniff (SSL Sniff) or Secure Sockets Layer Strip (SSL Strip), two methods not very easy to achieve, demonstrated and described by the researcher Moxie Marlinspike at the Black Hat Conference in 2009. SSL Sniff is a method by which an intruder intercepts a connection from the client´s side, generates and signs a certificate for the site it is connecting to. Then they pass that certificate chain to the client, make a normal SSL connection to the server, and pass data between client and server, decrypting and encrypting on each end. SSL Strip is a method of watching the traffic of http and https packets and convert them to log the packets and keep all relative data that go by. In the vast majority of publicly available Wi-Fi infrastructures, the cases that don’t use end-to-end encryption, the attacker doesn’t need to achieve a man-in-the-middle attack in order to capture the traffic between the two ends, because the communication is broadcasted and each and every user shares the same password within the Wi-Fi network.

Another way for a malicious user to manage the traffic within a Wi-Fi infrastructure is to create a rogue access point, known as rogue APS or rogue AP. A rogue APS is a Wi-Fi that has been installed within a secure network, but it is not authorized by the administration (“Rogue Access Point”, 2020). In the case of traffic capture, the rogue APS works as a clone of a legitimate Wi-Fi router, but with a stronger signal which enforces the user to connect to the rogue APS. Another way of creating a rogue APS is by having a totally free-of-password access point with an also stronger signal, creating tempting conditions for the unsuspecting user to login. At this point, it is worth mentioning that it is almost impossible for a malicious user to create a rogue access point if the keys/passwords given to each user are different for everyone.

5.6 Recommendations – Future Work

Although a lot of solutions have been provided by researchers in order to enhance Wi-Fi security (Xiong & Jamieson, 2013), very few can be applied in public available Wi-Fi networks for numerous reasons: firstly, the management/administration of a public place such as a restaurant or cafe will not consider the security of its customers as an issue; and secondly, even if they do, they will think over it twice, due to extra monetary or administrative effort to install, configure and maintain such a secure infrastructure. For some, a considerable solution would be the enforcement of strong password composition policies by the administration. But as it will become clearer further down this section, strong password composition policies are not a proper solution regarding shared passwords. Other solutions could come directly from Internet Service Providers (ISPs). ISPs should provide users/companies with Wi-Fi routers secured by-design, with a strong password security policy and traffic encryption.

An effective solution would be the adaptation of strong password composition policies from the administration/management (Narayanaswami & Raghunath, 2007), in order to enhance the security level of Wi-Fi networks. The necessity regarding the adaptation of strong password composition policies is more than crucial taking into consideration that, to the best of our knowledge, no router performs a check regarding password strength. A strong password composition policy should include a combination of uppercase and lowercase letters, numbers, and symbols, and typically be a minimum of fourteen characters long or more. Furthermore, other alternatives regarding the composition of strong passwords have been recently proposed, for example, “zxcvbn”, or a password strength estimator proposed by Daniel Lowe Wheeler (2016), which works by using leaked passwords and comparing its estimations (Egelman et al., 2013). Another alternative to enhance human-chosen text passwords has been proposed by Melicher et al. (2016), suggesting the usage of artificial neural networks to model text passwords’ resistance to guessing attacks and explore how different architectures and training methods impact neural networks’ guessing effectiveness. These might work well for a

corporate environment, where the Wi-Fi password would be a knowledge of the IT administration and only, but might not work for publicly available Wi-Fi networks, where the password is shared and given freely to everyone.

So as to avoid major security incidents, it is important not to use the same shared password for every user. Even if the password is 20 characters long and meets all the strong password policy requirements, once the password is known to everyone, then it loses its purpose of complexity and uniqueness. The administration of each public Wi-Fi infrastructure should include some form of endpoint encryption to prevent attacks like man-in-the-middle, TLS, and SSL, as described in Section 1.5. Apart from adopting end-to-end encryption solutions, such as SSL packet transfer, the most important thing in order to ensure up to a point the security of each user, is to use separate credentials for every single user, adopting an endpoint authentication strictly unique for everyone who wants to log in to a public Wi-Fi network.

Another solution, which acts mainly as an informative solution rather than a robust security alternative, is captive portals ("Man-in-the-middle attack", 2020). A captive portal may work as a landing page shown to users before they gain broader access to a URL or http-based Internet services, in this case, the Wi-Fi network. Administrators tend to adopt captive portals as a solution to well-known hotels and branded cafeterias (Hilton hotel, Starbucks café, etc.) to avoid any major problems since the users take responsibility for their actions.

In some other cases, the adaptation of login via Facebook or Google accounts to save time for the user, is possible. The company chooses Facebook/Google as the login platform in order to avoid the sharing of passwords ("Get Facebook Wi-Fi for Your Business", 2020). This solution also works as a captive-like-portal, but without the extra layer of endpoint authentication, it doesn't provide an add-on security layer.

A similar but more secure solution would be the provision of a secure connection between untrusted devices, based on a key exchange between the two parties as proposed by Ian Herwono and Paul W. Hodgson (2014). This solution will work in terms of a Virtual Private Network as it is described further below. Another alternative in a similar direction but less secure is the proposal of Lafuente et al. (2011), who suggest a secure

and trust-based Wi-Fi password sharing service. They proposed an architecture that would provide an alternative by which Wi-Fi network password sharing would be enabled by a social-networking oriented trust model approach. Each user of the network will be able to locate and connect to any available Wi-Fi network at any time.

Numerous solutions could be implemented also from the side of the ISPs (Bachy et al., 2015). A robust solution that would come directly from the ISP providers would be the provision to their customers with secure-by-design Wi-Fi routers that would enforce strong password composition policies (Lafuente et al., 2011) and will not include the administration credentials of each router on the back of the device. Again, this solution would be appropriate for home users that don't share their Wi-Fi credentials and for companies where the administration doesn't allow Wi-Fi sharing. But for a public Wi-Fi infrastructure, it will still be an incomplete solution. Another alternative regarding Internet Service Providers has to do with the adaptation of security practices that have to do with the Integrated Access Device (IAD) and its communications on the local loop (Shay et al., 2016). An IAD is a customer premises device that provides access to wide area networks and the Internet. Specifically, it aggregates multiple channels of information including voice and data across a single shared access link to a carrier or service provider PoP (Point of Presence) ("Integrated Access Device", 2020).

These solutions coming directly from the ISPs would enhance the level of security of the users enjoying these services by default. But, is this enough? In the introduction of the current chapter, it is mentioned that users tend to prefer publicly available Wi-Fi networks not only to reduce costs but also to save energy from the battery of their device. By using a more secure-by-design Wi-Fi router, the security will increase overall. But, would this suffice?

Enforcing security will not solve the problem entirely. When security issues arise, security awareness is one of the most effective ways to deal with the unauthorized access phenomenon. The main focus should be the training of the users in order to be aware of the threats and issues involved regarding the information and data transferred within the wireless connection of the public Wi-Fi network ("Integrated Access Device", 2020). Users should know that while connected to a Wi-Fi network, especially when it is

publicly available, man-in-the-middle attacks is a quite common situation especially for unprotected Wi-Fi infrastructures, which can be mitigated by adopting SSL in the between communication (A. Ochang et al., 2016), but when we use a shared password SSL will not be enough.

From the users' perspective, an appropriate suggestion is to rethink before logging in to an unknown Wi-Fi network. In case that a user will have to log in to an untrusted Wi-Fi network, then they should check that sharing network features are turned off, the firewall is enabled, HTTPS and SSL are used whenever it is possible, Wi-Fi is turned off when it is not in use. Additionally, an effective practice is considering using a virtual private network (VPN) ("Virtual Private Network", 2020). By using a VPN, a user is enabled to send and receive data across untrusted networks, as if the device that is used, it is directly connected to the private network. In any case, it is preferable to use cellular data rather than using public Wi-Fi networks from a security perspective.

The next step for future work regarding password analysis for publicly available passwords could be the analysis of one of the applications mentioned in Section 3 (Swift WiFi: Global Wi-Fi Sharing, 2016; Free WiFi Internet Finder, 2016; Free WiFi – Wiman, 2016; Show WiFi Password, 2016; WADA Wi-Fi Maps - Free Wi-Fi, 2016; WiFi Chua Free WiFi password, 2016; Wi-Fi Space, 2016) or in a combination of those that were not examined in a more detailed - database level in the current research.

5.7 Conclusions

By establishing sound data collection methodology and rigorously analyzing a large password corpus of publicly available Wi-Fi networks, hopefully, we have contributed to the emerging literature about this issue. The trickiest part of our research was the collection of the database and the analysis of the password corpus. Where possible, comparison to past analysis has been done as mentioned in the related work section. Wi-Fi network security brings numerous issues regarding the users, the administrator, and the Internet Service Provider. It is of high importance to promote solutions from both the

administration and the ISP, but regarding the single user, security is not an option unless it is enforced. Our analysis showed that publicly available Wi-Fi networks, even though they may administer critical data, have a significantly low level of security, regardless of continent, country, and city, due to the fact that credentials are shared with everyone who wants to use the publicly available Wi-Fi infrastructure. We have concluded some interesting results regarding users' behavior and parameters that may create the need for logging into a publicly available Wi-Fi network. The recommendations analyzed are not limited to those mentioned, but they showed the overall need for the enhancement of the security of public Wi-Fi networks. Even though we can observe the use of simple passwords in the vast majority of the existing public Wi-Fi infrastructures, alternatives were suggested in order to enhance security while using a public Wi-Fi network.

We expected to find significant differences in the level of security awareness between each city, according to different living standards, education, and many more similar attributes. However, security and privacy researchers have found that geographic location and demographic characteristics have little effect on security/privacy behaviors (Friedman et al., 2002). We expected to find more complex passwords in countries within the European Union or the United States of America where there is a higher level of security awareness as shown in the SANS report for 2016 (SANS Organization, 2017), compared with countries in North Africa or West Asia. Our initial hypotheses proved inconclusive, as we found that publicly available Wi-Fi networks play a one and only basic role worldwide regardless of geographic location and living standards, that is they help users to easily access the web.

Chapter 6: Epilogue

The contribution of this thesis is threefold. First, it addresses the multiple problems regarding the passwords of open Wi-Fi networks. Second, it covers the adaptation of security mechanisms regarding TDM technologies, and third, it thoroughly analyzes the implementation of IT infrastructures and websites with regard to cybersecurity and the GDPR.

This thesis analyzed for the first time a password database of more than one million passwords of publicly available Wi-Fi networks from the biggest cities and capitals around the world, employing an innovative methodology. In doing so, new insights were provided on how businesses who offer public Wi-Fi networks perceive security and ultimately deal with it. Furthermore, by comparing the database collected for this research with other private password datasets, this piece of work aims, for the first time in the literature, at laying down a more rounded idea of how convenience outweighs consequence, especially with how people use their mobile devices and, therefore, explaining security naivety. Our analysis showed that publicly available Wi-Fi networks, even though they may administer critical data, have a significantly low level of security, regardless of continent, country, and city, due to the fact that credentials are shared with everyone who wants to use the publicly available Wi-Fi infrastructure.

We expected to find significant differences in the level of security awareness between each city, according to different living standards, education, and other factors we examined. For example, we expected to find more complex passwords in countries within the European Union or the United States of America, where there is a higher level of security awareness as shown in the SANS report for 2016 (SANS Organization, 2017), compared with countries in North Africa or West Asia. Our initial hypotheses proved inconclusive. Indeed, security and privacy researchers have found that geographic location and demographic characteristics have little effect on security/privacy behaviors (Friedman et al., 2002). Therefore, we argued that publicly available Wi-Fi networks play a sole basic role worldwide regardless of geographic location and living standards: they help users access easily the web.

Moreover, TDM were explored as a technological option, focusing on the TDM deployed by the NLG and considerations for applied Internet Security solutions taking into account the GDPR requirements. The technological option of TDM adopted by large libraries has brought some fruitful outcomes regarding web-harvesting and web-archiving with the aim to collect, download, archive, and preserve content and works that are available on the Internet. TDM is used to index, analyze, evaluate, and interpret mass quantities of works including texts, sounds, images, or data through an automated "tracking and pulling" process of online material. Regardless of TDM's significant contribution to libraries worldwide, access to the web content and works available online are subject to restrictions by legislation, especially to laws pertaining to Copyright, Industrial Property Rights, and Data Privacy. For the first time, the benefits of TDM are analyzed as a technological option in conjunction with regulations on this matter. As far as Data Privacy is concerned, the application of the GDPR is considered an issue of vital importance for the smooth operation of TDM services offered by national libraries mostly in the EU Member States, which among other requirements mandates the adoption of privacy-by-design and advanced security techniques.

Lastly, the requirements of the GDPR, as well as their implementation with regard to applied Internet Security solutions for network infrastructures and websites were analyzed. While the Regulation offers a minimum set of technical Internet Security means to be taken into consideration by companies and organizations to achieve GDPR compliance, this thesis highlights the adaptation of strong security mechanisms that will not only set compliant infrastructures and websites with the GDPR but also maintain them strong and secure against most threats.

6.1 Publications

The contribution of this thesis can be further found in the following journals and peer-reviewed conference proceedings.

1) Journal Articles

a. **Vavousis, K.**, Papadopoulos, M., Polley, J., & Xenakis, C. (2020). A compliant and secure IT infrastructure for the National Library of Greece in consideration of internet security and GDPR. *Qualitative and Quantitative Methods in Libraries, 9(2)*, 219–236.

b. **Vavousis, K.**, Papadopoulos, M., Gerolimos, M., & Xenakis, C. (2020). Compliant and secure websites for the Greek Libraries Network of the National Library of Greece and each library-member of this Network in consideration of internet security and GDPR. *Qualitative and Quantitative Methods in Libraries, 9(3)*, 377–395.

c. Papadopoulos, M., Gerolimos, M., **Vavousis, K.**, & Xenakis, C. (2020). Text and Data Mining for the National Library of Greece in consideration of Internet Security and GDPR. *Qualitative and Quantitative Methods in Libraries, 9(3)*, 441–460.

d. Veroni, E., Ntantogian, C., **Vavousis, K.**, & Xenakis, C. (2021). *A Worldwide Empirical Analysis of Wi-Fi Passwords*. Manuscript submitted for publication.

References

A. Ochang, P., J. Irving, P., & O. Ofem, P. (2016). Research on Wireless Network Security Awareness of Average Users. *International Journal of Wireless and Microwave Technologies*, *6*(2), 21–29. DOI: 10.5815/ijwmt.2016.02.03

Adam. (October 12, 2015). What 10 million passwords reveal about the people who choose them. *NO2ID's Newsblog*. Retrieved from: https://www.no2id.net/newsblog/2015-10/what-10-million-passwords-reveal-about-the-people-who-choose-them/

Article 29 Working Party (April 10, 2014). Opinion 05/2014 on Anonymization Techniques *(*WP216*)*.

Announcing Our Worst Passwords of 2015. (2016). *Team Password*. Retrieved from: https://www.teamsid.com/worst-passwords-2015/

Bachy, Y., Nicomette, V., Alata, E., Kaaniche, M., & Courrege, J.-C. (August, 2015). Security of ISP Access Networks: Practical Experiments. In *2015 11th European Dependable Computing Conference (EDCC).* DOI: 10.1109/edcc.2015.27

Bajpai, P., Singh, N.R.A.J., & Singh, V. (2014). *Analysis of Current Wi-Fi Security Practices Via War*. *7*, 45–49.

Balasubramanian, N., Balasubramanian, A. & Venkataramani, A. (November 4-6, 2009). Energy consumption in mobile phones: a measurement study and implications for network applications. In *IMC '09 Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. Chicago, Illinois, USA.

Bashah Mat Ali, A., Yaseen Ibrahim Shakhatreh, A., Syazwan Abdullah, M., & Alostad, J. (2011). SQL-injection vulnerability scanning tool for automatic creation of SQL-injection attacks. *Procedia Computer Science*, 3, 453-458.

Bhiogade, M.S. (June 19-21, 2002). Secure Socket Layer. In *2002 Informing Science & IT Education Conference proceedings*. Cork, Ireland.

Blokdyk, G. (2019). *Mobile Device Management MDM – A Complete Guide*. (2019 Edition). 5STARCooks.

Blokdyk G. (2020). *Security Information and Event Management SIEM – A Complete Guide*. (2020 Edition). 5STARCooks.

Bonneau, J. (2012, April). The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*. DOI: 10.1109/SP.2012.49

Botti, M., Papadopoulos, M., Zambakolas, C., & Ganatsiou, P. (2019a). On the Eve of Web-Harvesting and Web-Archiving for Libraries in Greece. *Erasmus Law Review, 12*, 178-189.

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019b). Text and Data Mining in Directive 2019/790/EU. Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives. *Open Journal of Philosophy (OJPP), 9*, 369-395.

Callas, J., Donnerhacke, L., Finney, H., Shaw, D, & Thayer, R. (2007). OpenPGP Message Format. *IETF Proposed Standards Track*, (RFC 4880).

Caspers, M., Guibault, L., McNeice, K., Piperidis, S., Pouli, K., Eskevich, M., & Gavriilidou, M. (2016). Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments Using a Collaborative Knowledge and Open Information Approach. *Baseline Report of Policies and Barriers of TDM in Europe*. Retrieved from: https://cordis.europa.eu/project/id/665940/reporting/es

Cassola, A., Blass, E.-O., & Noubir G. (2015, October 12-16). Authenticating Privately over Public Wi-Fi Hotspots. In *CCS '15 Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp 1346-1357). Denver, Colorado, USA.Clement, J. (August 04, 2017). Global public Wi-Fi safety awareness 2017. *Statista*. Retrieved from https://www.statista.com/statistics/731556/worldwide-public-wifi-safety-awareness/

Cheng, N., Oscar Wang, X., Cheng, W., Mohapatra, P., & Seneviratne, A. (2013). Characterizing privacy leakage of public WiFi networks for users on travel. *Proceedings - IEEE INFOCOM, October 2015*, 2769–2777. DOI: 10.1109/INFCOM.2013.6567086

Chou, H.C., Lee, H.C., Hsueh, C.W., & Lai, F.P. (2012). Password cracking based on special keyboard patterns. *International Journal of Innovative Computing, Information and Control*, *8*(1 A), 387–402.

Chou, H.C., Lee, H.C., Yu, H.J., Lai, F.P., Huang, K.H., & Hsueh, C.W. (2013). Password cracking based on learned patterns from disclosed passwords. *International Journal of Innovative Computing, Information and Control*, *9*(2), 821–839.

Clement, J. (February 20, 2019). Public Wi-Fi usage of adults worldwide 2017. *Statista.* Retrieved from https://www.statista.com/statistics/731525/unsecured-wifi-risky-behavior-of-adults-global/

Conti, M., Dragoni, N., & Lesyk, V. (2016). A Survey of Man in the Middle Attacks.*IEEE Communications Surveys & Tutorials, 8(3)*, 2027-2051. DOI: 10.1109/COMST.2016.2548426

Dabrowski, A., Merzdovnik, G., Ullrich, J., Sandera, G., & Weippl, E. (2019). *Measuring Cookies and Web Privacy in a Post-GDPR World*. In D. Choffnes & M. Barcellos (eds), Passive and Active Measurement. PAM 2019. Lecture Notes in Computer Science (11419). Springer, Cham. DOI: 10.1007/978-3-030-15986-3_17

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (February 24-27, 2019). We Value Your Privacy… Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy.

In*2019 Network and Distributed System Security Symposium Proceedings*. San Diego, CA, USA. DOI: 10.14722/ndss.2019.23378

Dehalwar, V., Kalam, A., Lal Kolhe, M., & Zayegh, A. (December 21-23, 2017). Review of web-based information security treats in smart grid. In the *7th International Conference on Power Systems (pp.849-853)*. Pune, India. DOI: 10.1109/ICPES.2017.8387407

Desmedt, Y. (2005). Man-in-the-Middle Attack. In van Tilborg H.C.A., *Encyclopedia of Cryptography and Security*, (368). Springer. MA, USA: Springer.

DigiNinja, R. (2012). Pipal, Password Analyser. *DigiNinja Something about security*. Retrieved from: https://digi.ninja/projects/pipal.php

Directive 95/46/EC. *On the protection of individuals with regard to the processing of personal data and on the free movement of such data*. The European Parliament and Council. Retrieved at April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046

Directive (EU) 2019/790. *On copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.* The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/eli/dir/2019/790/oj

Directive 96/9/EC. *On the legal protection of databases*. The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

Directive 2001/29/EC. *On the harmonization of certain aspects of copyright and related rights in the information society*. The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32001L0029

Directive 2009/24/EC. *On the legal protection of computer programs* (Codified version). The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009L0024

Directive (EU) 2016/1148. *On measures for a high common level of security of network and information systems across the Union*. The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.194.01.0001.01.ENG

Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K., & Herley, C. (2013). Does my password go up to eleven? the impact of password meters on password selection. *Conference on Human Factors in Computing Systems - Proceedings*, 2379–2388. DOI: 10.1145/2470654.2481329

Eshete, B., Villafiorita, A., & Weldemariam, K. (August 22-26, 2011). Early Detection of Security Misconfiguration Vulnerabilities in Web Applications. *In the Sixth International Conference on Availability, Reliability and Security*. Vienna, Austria. DOI: 10.1109/ARES.2011.31

European Commission. (2018). *The GDPR: new opportunities, new obligations*, Luxembourg: Publications Office of the European Union.

European Data Protection Supervisor. (2018). Guidelines on the protection of personal data in IT governance and IT management of EU institutions. *European Data Protection Supervisor.* Retrieved on April 30, 2020 from: https://edps.europa.eu/sites/edp/files/publication/it_governance_management_en.pdf

Evans, N. J. (2009). *Information technology social engineering: an academic definition and study of social engineering-analyzing the human firewall* (Doctoral Dissertation). Iowa State University, Ames, IA, USA. Retrieved from: https://lib.dr.iastate.edu/cgi /viewcontent.cgi? article=1701&context=etd

Fahl, S., Harbach, M., Perl, H., Koetter, M., & Smith, M. (November 4-8, 2013). Rethinking SSL development in an appified world. *In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Berlin, Germany (pp.46-60). DOI: 10.1145/2508859.2516655

Free WiFi Internet Finder - Apps on Google Play. (2016). Retrieved from: https://play.google.com/store/apps/details?id=com.kapron.ap.hotspot

Free WiFi Winman - Apps on Google Play. (2016). Retrieved from: https://play.google.com/store/apps/details?id=me.wiman.androidApp

French, A., Guo, C., & Shim, J.P. (2014). Current Status, Issues, and Future of Bring Your Own Device (BYOD). *Communications of the Association for Information Systems, 35(10),* 191-197.

Friedman, B., Hurley, D., Howe, D.C., Felten, E., & Nissenbaum, H. (April 20-25, 2002). Users' Conceptions of Web Security: A Comparative Study. In *Conference on Human Factors in Computing Systems* (pp. 746–747). Minneapolis, MN, USA.

Greeklish. (2020). Retrieved from: https://en.wikipedia.org/wiki/Greeklish

Guessing entropy. (2020). Retrieved from: http://itlaw.wikia.com/wiki/Guessing_entropy

Gupta, S., &Gupta, B. (2015). Cross-Site Scripting (XSS) attacks and defense mechanisms: classification and state-of-the-art. *International Journal of System Assurance Engineering and Management, 8*, 512-530(2017). DOI: https://doi.org/10.1007/s13198-015-0376-0.

Hargreaves, I., Guibault, L., Handke, C., Martens, B., Lynch, R., & Filippov, S. (2014). Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining. *Report from the Expert Group, European Union*. Luxemburg: Publications Office of the European Union.

Hassan, M., Shamina Sultana, N., Marjan, A., Rafita, H., Fabiha Nawar, D., Mostafijur, R., Asif, S., & Hasan, S. (2018). *Broken Authentication and Session Management Vulnerability: A Case Study of Web Application*, 6.1-6.11. DOI: 10.5013/IJSSST.a.19.02.06

Hassan, M., Ali, M., Bhuiyan, T., Sharif, M., & Biswas, S. (October 18-20, 2018). Quantitative Assessment on Broken Access Control Vulnerability in Web Applications. In *International Conference on Cyber Security and Computer Science 2018*. Safranbolu, Karabuk, Turkey.

Hulton, D. (2002). Practical exploitation of RC4 weakness in WEP environments. In *HiverCon*.

IFLA. (2013). IFLA Statement on Text and Data Mining. *IFLA*. Retrieved from: https://www.ifla.org/publications/node/8225

Integrated access device. (2020). Retrieved from: https://en.wikipedia.org/wiki/Integrated_access_device

International Organization for Standardization/Technical Committee. (2008). *Health Informatics. Pseudonymization*. (ISO/TC 215). https://www.iso.org/committee/54960.html

Johnson, G., Shriver, S., & Goldberg, S. (2020). Privacy & Market Concentration: Intended & Unintended Consequences for the GDPR. *SSRN*. DOI: 10.2139/ssrn.3477686Get Facebook Wi-Fi for Your Business. (2020). Retrieved from: https://el-gr.facebook.com/business/facebook-wifi

Joshi, Y., Das, D., & Saha, S. (2009). Mitigating man in the middle attack over secure sockets layer. *2009 IEEE International Conference on Internet Multimedia Services Architecture and Applications, IMSAA 2009*. DOI: 10.1109/IMSAA.2009.5439461

Kantarcioglu, M., & Xi, B. (October 24-28, 2016). Adversarial Data Mining: Big Data Meets Cyber Security. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp.1866–1867*. Vienna, Austria.

Kelley, P. G., Komanduri, S., Mazurek, M. L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L. F., & López, J. (2012). Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. *Proceedings - IEEE Symposium on Security and Privacy*, 523–537. DOI: 10.1109/SP.2012.38

Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., & Egelman, S. (2011). *Of passwords and people*. 2595. DOI: 10.1145/1978942.1979321

Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced social engineering attacks. *Journal of*

Information Security and Applications, *22*, 113–122. DOI: 10.1016/j.jisa.2014.09.005

Lafuente, C. B., Titi, X., & Seigneur, J. M. (2011). Flexible communication: A secure and trust-based free Wi-Fi password sharing service. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing* http://www.europarl.europa.eu/supporting-analyses

Kuner, C., Bygrave, L., Docksey, C., & Drechsler, L. (2020). *The EU General Data Protection Regulation (GDPR) – A Commentary.* Oxford/UK: Oxford University Press.

Lafuente, C. B., Titi, X., & Seigneur, J. M. (2011). Flexible communication: A secure and trust-based free Wi-Fi password sharing service. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications, TrustCom 2011, 8th IEEE Int. Conf. on Embedded Software and Systems, ICESS 2011, 6th Int. Conf. on FCST 2011, December 2014*, 706–713. DOI: 10.1109/TrustCom.2011.91

Leite, G.S., & Albuquerque, A.B. (2019). *An Approach for Reduce Vulnerabilities in Web Information Systems*. In R. Silhavy, P. Silhavy & Z. Prokopova Z. (eds). *Intelligent Systems in Cybernetics and Automation Control Theory*. *CoMeSySo 2018. Advances in Intelligent Systems and Computing (860)*. Springer, Cham. DOI: 10.1007/978-3-030-00184-1_9

Lindèn, T., Khandelwal, R., Harkous, H., & Fawaz, K. (2020). The Privacy Policy Landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies*, *2020(1)*, 47-64. DOI: https://doi.org/10.2478/popets-2020-0004

Lindèn, T. (2019). *Building a secure WordPress website with plugins* (Bachelor's Thesis)., LAMK, Lahti, Finland. Retrieved on August 12, 2020 from: https://www.theseus.fi/bitstream/handle/10024/263175/Tuomas_Lind%c3%a9n.pdf?sequence=2&isAllowed=y

LinkedIn Revisited - Full 2012 Hash Dump Analysis. (2016, May 19). *Korelogic Security Blog*. Retrieved from: https://blog.korelogic.com/blog/2016/05/19/linkedin_passwords_2016

Lunden, I. (2012, July 12). Yahoo Confirms, Apologizes For The Email Hack, Says Still Fixing. Plus, Check If You Were Impacted (Non-Yahoo Accounts Apply). *TechCrunch*. Retrieved from: https://techcrunch.com/2012/07/12/yahoo-confirms-apologizes-for-the-email-hack-says-still-fixing-plus-check-if-you-were-impacted-non-yahoo-accounts-apply/

Man-in-the-middle attack. (2020). Retrieved from: https://en.wikipedia.org/wiki/Man-in-the-middle_attack

Marlinspike, M. (2009). More tricks for defeating SSL in practice. In *Black Hat Conference*. Retrieved January 10, 2020. Retrieved from: http://www.blackhat.com/presentations/bh-usa-09/MARLINSPIKE/BHUSA09-Marlinspike-DefeatSSL-SLIDES.pdf

Melicher, W., Ur, B., Segreti, S. M., Komanduri, S., Bauer, L., Christin, N., & Cranor, L. F. (2016). Fast, lean, and accurate: Modeling password guessability using neural networks. *Proceedings of the 25th USENIX Security Symposium*, 175–191.

Microsoft Corporation. (2019). *Email encryption in Office 365.* Retrieved on April 30, 2020 from: https://docs.microsoft.com/en-us/microsoft-365/compliance/email-encryption

Most Expensive Countries for International Roaming. (n.d.). *WORLDSIM*. Retrieved from: https://www.worldsim.com/blog/expensive-countries-international-roaming? store=eu

Narayanaswami, C., & Raghunath, T. (2007). *(12) United States Patent*. *2*(12).

NTRUEncrypt. (2020). Retrieved from: https://en.wikipedia.org/wiki/NTRUEncrypt

Osincev, A., & Laponina, O. (2019). Vulnerability Testing in Web Applications External Entities XML. *International Journal of Open Information Technologies, 7(10)*, 71-79.

Paaβ, G., Reinhardt, W., Püping, S., & Wrobel, S. (2014). Data Mining for Security and Crime Detection. In C.S. Gal, P.B. Kantor, & B. Shapira (Eds.), *NATO Science for Peace and Security Series, D: Information and Communication Security* (Vol. 15). (pp.56-70). Retrieved from: http://ebooks.iospress.nl/volumearticle/23568

Packet Capture - Apps on Google Play. (2016). Retrieved from: https://play.google.com/store/apps/details?id=app.greyshirts.sslcapture

Papadopoulos, M. (2020). Scientific Research, Web Harvesting and Text & Data Mining. *In* V. Zachou, *Archives and Cultural Readings* (130-148). Athens: Ocelotos Publications.

Papadopoulos, M., Botti, M., Ganatsiou, P., & Zampakolas, C. (2020). Emporical Research on Web Harvesting in the process of Text and Data Mining in National Libraries of EU Member States, *Open Journal of Philosophy (OJPP), 10*, 369-395.

Patel, S., Rathod, V.R., & Parikh, S. (December 8-9, 2011). Joomla, Drupal and WordPress – a statistical comparison of open source CMS. In the 3rd International Conference on Trendz in Information Sciences & Computing. Chennai, India. DOI: 10.1109/TISC.2011.6169111

Personal data pseudonymization: GDPR pseudonymization what and how. (n.d.). Retrieved on April 30, 2020 from: https://www.i-scoop.eu/gdpr/pseudonymization

Politou, E., Michota, A., Alepis, E., Pocs, M., & Patsakis, C. (2018). Backups and the right to be forgotten in the GDPR: An uneasy relationship. *Computer Law & Security Review, 34(6)*, 1247-1257. DOI: 10.1016/j.clsr.2018.08.006

Ponemon Institute LLC. (2017). 2017 Cost of Data Breach Study, Global Overview, Benchmark research sponsored by IBM Security. Retrieved on April 30, 2020 from: https://www.ibm.com/security/data-breach

Ponemon Institute LLC. (2019). Cost of Data Breach Report, Global Overview, Benchmark research sponsored by IBM Security. Retrieved on April 30, 2020 from: https://www.ibm.com/security/data-breach

Prince, M. (2018, August 23). The Relative Cost of Bandwidth Around the World. *The Cloudflare Blog*. Retrieved from: https://blog.cloudflare.com/the-relative-cost-of-bandwidth-around-the-world/

Private information retrieval. (2020). Retrieved from https://en.wikipedia.org/wiki/Private_information_retrieval

Raju, L.K., & Nair, R. (2015, November). Secure Hotspot a novel approach to secure public Wi-Fi hotspot. In *Control Communication & Computing India (ICCC), 2015 International Conference* (pp. 642-646). Trivandrum, India. DOI: 10.1109/ICCC.2015.7432975

Regulation (EU) 2016/679. *On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC* (General Data Protection Regulation). The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Regulation (EU) 2019/881. *On ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013* (Cybersecurity Act). The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/eli/reg/2019/881/oj

Regulation (EU) No 526/2013. *Concerning the European Union Agency for Network and Information Security (ENISA) and repealing Regulation (EC) No 460/2004.* The European Parliament and Council. Retrieved on April 30, 2020 from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R0526 Rogue access point. (2020). Retrieved from: https://en.wikipedia.org/wiki/Rogue_access_point

Sanchez-Rola, I., Dell' Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vernier, P-A., & Santos, I. (2019). Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. *In Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. Auckland, New Zealand. DOI: https://doi.org/10.1145/3321705.3329806

SANS Organization. (2017). 2016 Security Awareness Report. *SANS Security Awareness*. Retrieved from: https://www.sans.org/security-awareness-training/reports/2016-security-awareness-report

Scheeres, J. W. (2012). *Establishing the Human Firewall: Reducing an Individual's Vulnerability to Social Engineering Attacks*. BiblioScholar.

Schweitzer, D., Boleng, J., Hughes, C., & Murphy, L. (2009). Visualizing Keyboard Pattern Passwords. *6th International Workshop on Visualization for Cyber Security 2009, VizSec 2009 - Proceedings*, 69–73. DOI: 10.1109/VIZSEC.2009.5375544

Shannon entropy. (2020). Retrieved from: https://en.wiktionary.org/wiki/Shannon_entropy

Shay, R., Komanduri, S., Durity, A. L., Huh, P., Mazurek, M. L., Segreti, S. M., Ur, B., Bauer, L., Christin, N., & Cranor, L. F. (2016). Designing password policies for strength and usability. *ACM Transactions on Information and System Security*, *18*(4). DOI: 10.1145/2891411

Shen, C., Yu, T., Xu, H., Yang, G., & Guan, X. (2016). User practice in password security: An empirical study of real-life passwords in the wild. *Computers and Security*, *61*, 130–141. DOI: 10.1016/j.cose.2016.05.007

Show Wifi Password - Root for Android - APK Download. (2016, July 12). Retrieved from: https://apkpure.com/show-wifi-password-2016-root/show.wifi.password.hienthi.matkhau.wifi.chua.ketnoi

Suh-Lee C., Jo, J-Y., & Kim, Y. (October 17-19, 2016). Text mining for security threat detection discovering hidden information in unstructured log messages. *2016 IEEE Conference on Communications and Network Security (CNS).* Philadelphia, PA, USA.

Swift WiFi - Free WiFi Hotspot Portable - Apps on Google Play. (2016). Retrieved from: https://play.google.com/store/apps/details?id=mobi.wifi.toolbox

Symantec Corporation. (2019). Encryption Solutions for Email Powered by PGP™ Technology, 21276730-9 02/17. Retrieved on April 30, 2020 from: https://www.symantec.com/content/dam/symantec/docs/data-sheets/encryption-solutions-for-email-en.pdf

Tankovska, H. (August 27, 2020). Global public Wi-Fi hotspots 2016-2022. *Statista*. Retrieved from https://www.statista.com/statistics/677108/global-public-wi-fi-hotspots/

Transport Layer Security. (2020). Retrieved from: https://en.wikipedia.org/wiki/Transport_Layer_Security

Troy Hunt. (2011, June 05). A brief Sony password analysis. *Troy Hunt*. Retrieved from: https://www.troyhunt.com/brief-sony-password-analysis/

Vavousis, K., Papadopoulos, M., Polley, J., & Xenakis, C. (2020). A compliant and secure IT infrastructure for the National Library of Greece in consideration of internet security and GDPR. *Qualitative and Quantitative Methods in Libraries, 9(2)*, 219–236.

Ved, A. (July 24, 2017). Privacy and Security by design is a crucial step for privacy protection. *Least Authority*. Retrieved on April 30, 2020 from: https://leastauthority.com/blog/privacy-and-security-by-design-is-a-crucial-step-for-privacy-protection/

Virtual private network. (2020). Retrieved from: https://en.wikipedia.org/wiki/Virtual_private_network

WADA Wi-Fi Maps – Free Wifi. (2016). Retrieved from: https://play.google.com/store/apps/details?id=com.wada.wifil

Ward C., & Pritam, N. (2017). Cyberespionage and ransomware attacks are on the increase warns the Verizon 2017 Data Breach Investigations Report. *Verizon*. Retrieved on April 30, 2020 from: https://www.verizon.com/about/news/cyberespionage-and-ransomware-attacks-are-increase-warns-verizon-2017-data-breach

Wardialing. (2020). Retrieved from https://en.wikipedia.org/wiki/Wardialing

Wheeler, D. L. (August 10 - 12, 2016). ZxCVBN: Low-budget password strength estimation. *Proceedings of -the 25th USENIX Security Symposium*, 157–173. Austin, TX, USA.

WiFi Chua – Free WiFi password. (2016). Retrieved from: https://play.google.com/store/apps/details?id=com.bangdev.wifichua

WiFi Map. (2016). #1 WiFi Finder. Retrieved from: http://www.wifimap.io/

Wi-Fi Protected Access. (2020). Retrieved from: https://en.wikipedia.org/wiki/Wi-Fi_Protected_Access

Wi-Fi Space. (2016). Retrieved from: http://wifispc.com/mobile-apps.html

Wired Equivalent Privacy. (2020). Retrieved from: https://en.wikipedia.org/wiki/Wired_Equivalent_Privacy

Wool, A. (April 27, 2010). Trends in Firewall Configuration Errors: Measuring the Holes in Swiss Cheese. *IEEE Internet and Computing, 14(4)*, 58-65.

Xiong, J. & Jamieson, K. (2013). SecureArray: improving wifi security with fine-grained physical-layer information. In *Proceedings of the 19th annual international conference on Mobile computing & networking* (pp. 441-452). ACM, New York, NY, USA.

Zimmermann, P. (1995). *The Official PGP User's Guide.* Cambridge. MA, USA: MIT Press.

Zuiderveen Borgesius, F., Kruikemeier, S., Boerman, S., & Helberger, N. (2017). Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation. *European Data Protection Law Review, 3(3),* 353-368.

3 Phases of protection by a Data Leakage Prevention (DLP) plan. (April 3, 2017). *Prov International*. Retrieved on April 30, 2020 from: https://www.provintl.com/blog/3-phases-of-protection-by-a-data-leakage-prevention-dlp-plan,

Additional bibliography

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019c). Text and Data Mining in the EU 'Acquis Communautaire'. Tinkering with TDM & Digital Legal Deposit. *Erasmus Law Review, 12*, 190-208.

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019d). Legal and Technical Issues for Text and Data Mining in Greece. *In the Proceedings of Computer Ethics—Philosophical Enquiry (CEPE) Conference, pp 19*. Norfolk, Virginia, USA. DOI: 10.25884/yp3n-dq78

De Wolf & Partners. (2014). *Study on the Legal Framework of Text and Data Mining (TDM).* Publication Office of the EU.

European Commission & COM 192 Final. (2015). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions: A Digital Single Market Strategy for Europe* (Document: 52015DC0192). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192

European Commission & SWD 301 Final Part 1/3. (2016). *Commission Staff Working Document: Impact Assessment on the Modernization of EU Copyright Rules* (Document: 52016SC0302). Retrieved from: https://ec.europa.eu/digital-single-market/en/news/impact-assessment-modernisation-eu-copyright-rules

European Commission & SWD 301 Final Part 2/3. (2016). *Commission Staff Working Document: Impact Assessment on the Modernization of EU Copyright Rules* (Document: 52016SC0302). Retrieved from: https://ec.europa.eu/digital-single-market/en/news/impact-assessment-modernisation-eu-copyright-rules

European Commission & SWD 302 Final Part 3/3. (2016). *Commission Staff Working Document*: *Executive Summary of the Impact Assessment, on the Modernization of EU Copyright Rules* (Document: 52016SC0302). Retrieved from: https://ec.europa.eu/digital-single-market/en/news/impact-assessment-modernisation-eu-copyright-rules

European Copyright Society. (January 24, 2017). General Opinion on the EU Copyright Reform Package. *Instituut Voor Informatierecht / Institute for Information Law NL*. Retrieved from: https://www.ivir.nl/publicaties/download/ECS_opinion_on_EU_copyright_reform.pdf

Feiler, L., Forgó, N., & Weigl, M. (2018). *The EU General Data Protection Regulation (GDPR): A Commentary*. German Law Publishers. Frankfurt, Germany.

Geiger, C., Frosio, G., & Bulayenko, O. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-Legal Aspects. Retrieved from:

Linder, A. (2016). *European Data Protection Law – General Data Protection Regulation 2016.* CreateSpace Independent Publishing Platform.

Markham, K. (2020). *A Practical Guide to the General Data Protection Regulation (GDPR): 2nd Edition.* Somerset/UK: Law Brief Publishing.

Sag, M. (2019). The New Legal Landscape for Text Mining and Machine Learning. *Journal of the Copyright Society of the USA, 66,* 291-365.

Voigt, P., & Brussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. New York/USA: Springer Publishing.