UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

EPOS Security & GDPR Compliance

Fábio André de Sousa Pereira

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Paul Crocker
Co-orientador: Prof. Doutor Valderi R. Q. Leithardt

Covilhã, setembro de 2020

Acknowledgements

I would like to thank Professor Dr. Paul Andrew Crocker and Professor Dr. Valderi R. Q. Leithardt for their supervision. I would also like to thank my family. I would like to thank *SEGAL* and all its students and professors, who, in one way or another, have supported me throughout the years and also The University of Beira Interior and the Institute of Telecommunications.
This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/EEA/50008/2020 and also by the EPOS-IP European Union Horizon 2020 research and innovation program under grant agreement Nº 676564.

Resumo

Desde maio de 2018 que as empresas precisam de cumprir o Regulamento Geral de Proteção de Dados (GDPR). Isso significa que muitas empresas tiveram que mudar seus métodos de como recolhem e processam os dados dos cidadãos da UE. O processo de conformidade pode ser muito caro, por exemplo, são necessários recursos humanos mais especializados, que precisam estudar os regulamentos e depois implementar as alterações nos aplicativos e infraestruturas de TI. Com isso novas medidas e métodos precisam ser desenvolvidos e implementados, tornando esse processo caro.

Este projeto está inserido no projeto European Plate Observing System (EPOS). O EPOS permite que dados sobre ciências da terra de vários institutos de pesquisa na Europa sejam compartilhados e usados. Os dados são armazenados em base de dados e em alguns sistema de ficheiros e além disso, existem *web services* para controle e mineração de dados. O projeto EPOS é um sistema distribuído complexo e portanto, é importante garantir não apenas sua segurança, mas também que seja compatível com o GDPR. Foi identificada a necessidade de automatizar e facilitar esse processo, em particular a necessidade de desenvolver uma ferramenta capaz de analisar aplicações *web*. Essa ferramenta, chamada PrivAcy, Data REgulation and Security (PADRES) pode fornecer às empresas uma maneira mais fácil e rápida de verificar o grau de conformidade com o GDPR com o objetivo de avaliar e implementar quaisquer alterações necessárias.

Com isto, esta ferramenta contém os pontos principais do General Data Protection Regulation (GDPR) organizado por princípios em forma duma *lista de verificação*, os quais são respondidos manualmente. Como os conceitos de privacidade e segurança se complementam, foi também incluída a procura por vulnerabilidades em aplicações web. Ao integrar as ferramentas de código aberto como o Network Mapper (NMAP) ou Zed Attack Proxy (ZAP), é possível então testar a aplicações contra as vulnerabilidades mais frequentes segundo o Open Web Application Security Project (OWASP) Top 10.

Aplicando esta ferramenta no EPOS, a maioria dos pontos relativos ao GDPR foram respondidos como estando em conformidade apesar de nos restantes terem sido geradas as respetivas sugestões para ajudar a melhorar o nível de conformidade e também melhorar o gerenciamento geral dos dados. Na exploração das vulnerabilidades foram encontradas algumas classificadas com risco elevado mas na maioria foram encontradas mais com classificação média.

Palavras-chave

GDPR, Vulnerabilidades, Segurança, Conformidade

Resumo alargado

Nos dias de hoje, os dados podem ser extraídos de praticamente todo o lado, desde roupas com pequenos *wearables* até ao histórico de navegação, dados estão presentes em todos os lugares. Tornou-se então, um ativo muito útil e importante para as empresas e isso significa apenas uma coisa, conhecimento. Esse dá então a capacidade às empresas de segmentar anúncios para pessoas específicas de acordo com seus gostos, de ter a inteligência por trás dos sistemas de recomendações encontrados em serviços de streaming, como o Netflix ou o Spotify, ou a capacidade de estudar ritmos cardíacos irregulares, encontrados no Apple Watch, torna este recurso num dos mais valiosos à sua disposição, permitindo-lhes ganhar muito dinheiro. Também permite o desenvolvimento de aplicações e ferramentas que facilitam nossas vidas, de uma maneira que não nos importamos em fornecer os nossos dados pessoais, de forma a ter serviços personalizados para nós. Essa interação entre humanos e computadores continua a aumentar e tornou-se tão natural para nós que parece quase ”invisível”. Esse conceito é definido pelo termo UbiComp, que significa computação ubíqua. Esse termo foi introduzido por Mark Weiser em seu artigo ”The Computer for the 21st Century” [1], que começa por dizer que ”The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it”. Esta citação, expressa perfeitamente o que é um sistema UbiComp, fazendo com que nós já não consigamos viver sem o auxílio deles. Estes sistemas, por estarem diretamente conectados ao nosso modo de vida, recolhem e processam os nossos dados, alguns deles classificados como privados. Como parece uma boa troca fornecer nossos dados, muitas vezes livremente, para receber serviços personalizados, um problema surge quando nossa privacidade é violada, e com isso, informações pessoais que não deviam estar disponíveis para outros, podem ser acessadas por qualquer pessoa com boas ou más intenções. De forma a proteger e regular a privacidade dos dados dos cidadãos da Europe Union (EU), desde maio de 2018 tornou-se obrigatório o cumprimento do Regulamento Geral de Proteção de Dados. O GDPR procura oferecer aos proprietários dos dados a possibilidade de controlar e proteger seus próprios dados. As empresas que não estiverem em conformidade com o GDPR podem pagar multas até 20 milhões de euros ou 4% do seu faturamento anual [10]. Como cada pequena parte da tecnologia compartilha seus dados com outros dispositivos, também se tornou extremamente importante garantir que, não apenas as comunicações são seguras, mas também o local onde as informações são guardadas precisa de ser virtualmente e fisicamente seguro.

Antes do GDPR ter sido adotado em 2016, já existia em actividade uma diretiva desde 1995, chamada Proteção Europeia de Dados. Esta diretiva foi antecessor do GDPR e foi criada para estar em conformidade com o artigo 8 da Carta dos Direitos Fundamentais da União Europeia [2]. O objetivo era criar uma estrutura que pudesse garantir a segurança e a liberdade dos dados pessoais dos indivíduos em todos os países da EU e também servir como orientação sobre como os dados devem ser armazenados, processados e transmitidos [3]. Com os avanços da tecnologia e da globalização, surgiram novos desafios em relação à proteção de dados [14]. Esses desafios exigiram o desenvolvimento de uma nova abordagem que pudesse garantir o direito de proteção de dados para indivíduos e seus dados pessoais. O conceito de dados pessoais é importante definir, porque é um dos principais conceitos para a proteção dos indivíduos [14]. Dados pessoais são ”quaisquer dados que possam ser vinculados a uma pessoa específica” [2], sendo esses associados direta ou indiretamente. Em qualquer dado, podemos ter identificadores pessoais, como o nome completo, número de identificação nacional ou identificadores indiretos,

como o endereço Protocolo da Internet (IP) ou fotos. Se os dados não possuem nenhum desses identificadores, esses dados são chamados anónimos [2] e, nesse caso, o GDPR não se aplica [15]. Mesmo que os dados sejam anónimos, eles podem ser identificados novamente usando uma técnica chamada “deanonimização”. Outro conceito importante do GDPR são os papéis dos intervenientes. Atualmente, existem três definidos, sendo o primeiro o titular dos dados. Este é a pessoa cujos dados pessoais vão ser recolhidos ou processados. Como o GDPR visa proteger esses indivíduos, eles têm direitos que vão do GDPR artigo 12 ao artigo 23 [4]. Após a recolha dos dados, o responsável pelo tratamento assume a responsabilidade, o que, de acordo com o Artigo 4 do EU GDPR, significa que ele é a pessoa que determina os objetivos do processamento. Finalmente, esse é feito pelo processador de dados em nome do controlador. Por exemplo, um supermercado é o controlador de dados, tendo a função de recolher os dados dos seus clientes quando estes compram alguns artigos. Em seguida, outra organização tendo o papel de processador, armazena e processa os dados fornecidos pelo controlador. Tanto o controlador quanto o processador são responsáveis pelo tratamento dos dados pessoais.

Para o processamento de dados, o GDPR define um conjunto de sete princípios [5], que são os *Legalidade*, *Justiça e Transparência*, *Limitação de Propósitos*, *Minimização de Dados*, *Precisão*, *Limitação de Armazenamento*, *Integridade e Confidencialidade* e *Accountability*.

Os campos de estudo, GDPR e segurança, complementam-se. Não há empresa em conformidade com o GDPR se os seus dados não estiverem seguros. No entanto, o oposto não funciona da mesma maneira. Os dados podem se encontrar seguros sem estarem em conformidade. Para estarem em conformidade, os dados devem estar seguros.

Com isto, o GDPR tornou-se uma prioridade para as organizações, mas também se tornou um problema, porque alguns deles não estão preparados para as mudanças que precisam ser feitas e não estão cientes das conseqüências que esse descumprimento pode trazer para eles. Estudos descobriram que esses problemas ocorrem devido ao fato de a regulamentação atual ser ”vaga, ambígua e detalhada”, o que significa que qualquer pessoa que não possua a proficiência necessária pode encontrar algumas dificuldades para entender a regulamentação. Por exemplo, a lei GDPR diz que as empresas devem fornecer um nível razoável de proteção de dados pessoais [6], mas a palavra ”razoável” não está bem definida. Também a promoção da ”privacidade por design”, sem haver um guia adequado de como isso pode ser alcançado. Este conhecimento que precisa de ser obtido, geralmente é um processo dispendioso, principalmente para pequenas e médias empresa que não possuem um departamento jurídico ou não podem contratar assessoria jurídica. Além disso, há também uma barreira para na altura de fazer os requisitos, em engenharia de software, no acesso à conformidade legal, derivada de dois grandes problemas [7]. O primeiro diz respeito a determinar quais regulamentos podem ser aplicados e o segundo está relacionado com a capacidade de desenvolver as políticas necessárias, que envolvem esses regulamentos. Mesmo depois de estabelecer esses requisitos, extrair informação dos regulamentos pode ser um trabalho propenso a erros.

Complementado o GDPR e indo ao encontro dos objectivos definidos, para analisar as vulnerabilidades encontradas, seguiu-se o ranking OWASP top 10 [8]. Como o nome sugere, é uma compilação dos dez riscos mais críticos em aplicações web. Isso é possível devido aos envios de dados recolhidos por empresas especializadas em segurança e, em seguida, os itens da lista são selecionados e ordenados de forma decrescente, em relação à combinação com ”estimativas consensuais de exploração, detectabilidade e impacto” [9]. Recomenda-se que este relatório seja incorporado nos relatórios de segurança da empresa para minimizar e mitigar os riscos de segurança. [10].

Além do OWASP, tornou-se uma prática atual o uso de aplicações Open Source software (OSS) em

execução no Linux Operating System (OS). Essa escolha pode ser justificada, porque o código-fonte do OSS será exposto a avaliações independentes, aumentando a probabilidade de haver correções de bugs e também um ponto muito importante é que geralmente é gratuito. Portanto, usar OSS para procurar vulnerabilidades pode ser realmente importante, porque provavelmente as encontrará, enquanto o uso de software proprietário pode não identificá-las intencionalmente.

De forma a conseguir testar e validar o software desenvolvido, foi estudado o sistema EPOS. Este envolve o estudo das ciências da Terra composto por diversos assuntos, como geologia, sismologia e geodésia. Esse estudo é conseguido devido à partilha e uso de dados sobre ciências da Terra de vários institutos de pesquisa na Europa, com o objectivo de serem monitorados para que possamos entender melhor o dinâmico e complexo sistema da Terra.

O Sistema Europeu de Observação de Placas, EPOS ", é um plano a longo prazo, para facilitar o uso integrado de dados, produtos de dados e infraestruturas de pesquisa distribuídas para ciências da Terra sólidas na Europa" [11]. Para fazer essa integração, existe um software chamado Geodetic Linking Advanced Software System (GLASS), Sistema de Software Avançado de Ligação Geodésica, responsável por esse processo.

Sendo o EPOS um sistema complexo e distribuído é importante garantir não apenas a sua segurança, mas também a sua conformidade com GDPR. Foi identificada a necessidade de automatizar e facilitar esse processo, em particular a necessidade de desenvolver uma ferramenta e uma metodologia de fluxo de trabalho que possam analisar aplicações Web. Essa ferramenta pode fornecer às empresas uma maneira mais fácil e rápida de verificar o grau de conformidade com o GDPR, a fim de implementar as alterações necessárias.

Com isto, esta ferramenta contém os pontos principais do GDPR organizado por princípios em forma de *checklist*, os quais são respondidos manualmente. Como os conceitos de privacidade e segurança se complementam, foi também incluída a procura de vulnerabilidades em aplicações web. Ao integrar as ferramentas de código aberto, como o NMAP ou ZAP, é possível então testar a aplicação contra as vulnerabilidades mais frequentes segundo o OWASP Top 10. Na fase final, o objetivo é poder exportar essa ferramenta para que possa ser usada em outras plataformas. Considerou-se que na primeira abordagem ao problema analisado, a ferramenta a desenvolver está apenas disponível localmente, de modo que o utilizador final só precise de a instalar e interagir com a mesma através do browser, principalmente porque esta se trata de uma análise a um sistema e também porque essa solução se destina somente para demonstração, não sendo assim necessário estar disponível on-line, que adicionaria consequentemente mais complexidade à solução. Respondendo a isso, a solução é baseada em um modelo cliente-servidor, em que o cliente é o front-end e o servidor o back-end. O back-end foi desenvolvido com base na arquitetura Representational State Transfer (REST), usando uma Application Programming Interface (API) para se comunicar com serviços. Como a solução também envolve o uso de um banco de dados para fornecer dados e também armazená-los, foi usado o SQLite para fornecer essas funcionalidades, porque é uma base de dados pequena e rápido, encaixando-se exatamente na solução proposta.

A interface do front end visa oferecer ao utilizador final uma visão dos princípios e dos pontos de regulação correspondentes. Além disso, com base na análise realizada, também é fornecida uma secção para verificar o histórico dos relatórios.

Quando o usuário decide fazer uma nova revisão de conformidade, ele deve escolher em qual software e em que país a revisão deve se basear. Este é o primeiro passo, a escolha do software precisa ser feita, porque uma empresa pode ter vários softwares e, mesmo que sejam usados juntos, como estudado anteriormente, é mais fácil analisar quando eles estão divididos. A

necessidade de escolher o país vem apoiar aqueles países que possuem outras leis além do GDPR que precisam ser cumpridas.

Para otimizar a distribuição e a execução da solução, foi usado o Docker para fornecer essas vantagens. Este é uma plataforma que permite uma maneira mais fácil de criar, implementar e executar aplicativos usando containers. Permite apenas interagir com containers únicos, portanto, um dos problemas ocorre quando há vários containers a serem geridos. Para resolver isso, é possível usar *Compose*, permitindo definir e executar múltiplos containers. Isso é obtido através de um documento chamado *docker-compose.yaml*, que engloba a definição de cada *Dockerfile* e também oferece a possibilidade de ter seus próprios volumes e configurações de rede. Os testes realizados para o EPOS foram relativamente diretos e realizados pelo autor desta dissertação, uma vez que ele também era membro deste projeto e estava diretamente envolvido no seu desenvolvimento.

Como explicado anteriormente, dentro do EPOS existem vários portais de dados, disponíveis em uma landing page. Portanto, primeiro a landing page será avaliada e, para entender a profundidade das ferramentas de avaliação de segurança, cada uma das sub páginas será avaliada e, no final, comparada entre elas. Além disso, como algumas páginas usam dados pessoais de forma independente, as perguntas relativas ao GDPR também será respondida.

Ao executar a solução com a Uniform Resource Locator (URL) que serve de entrada para os *softwares* do EPOS [12] , as perguntas GDPR são deixadas sem resposta, por não recolherem dados pessoais, não se apliacando assim o GDPR. Ao enviar então o formulário, foram marcadas as caixas para executar NMAP, OWASP ZAP, Wapiti e um scanner de cookies.

Analisando o relatório completo gerado após a conclusão de todas as verificações, vêm as informações relacionadas aos cookies encontrados reportando que não foram encontradas nenhumas a ser usadas. Observando os devtools disponíveis no browser, é possível dizer que as informações fornecidas estão corretas.

No relatório do NMAP constatou-se que 948 portos encontram-se fechados enviando uma resposta de reset, normalmente associado a respostas de portos fechados. Além disso, foram encontradas 10 portos abertos, identificando o serviço em execução em cada uma e, como foi fornecido à opção de script NMAP, ele também procurou por vulnerabilidades em cada serviço. Alguns dos serviços encontrados foi o do porto 21 que tem o serviço File Transfer Protocol (FTP) em execução, representando alguns problemas, pois as autenticações usam senhas em texto sem formatação, portanto, qualquer ataque Man In The Middle (MITM) pode ver as informações sendo trocadas e, em seguida, comprometer o sistema. Portanto, para superar esse problema, é recomendável usar SSH File Transfer Protocol (SFTP). A porta 22 também está aberta, aceitando conexões ssh, usando o serviço OpenSSH, na versão 5.3 e usando o protocolo 2.0, que oferece autenticação usando algoritmos de hash SHA-2 em vez do SHA-1 usado na versão anterior ou autenticação usando FIDO /U2F [13]. Além disso, está incluída uma lista do Common Vulnerabilities and Exposures (CVE), que é uma lista de entradas que contêm vulnerabilidades de segurança conhecidas publicamente, para esse serviço na versão 5.3. Com base nisso, é possível seguir o link disponível ou pesquisando em qualquer mecanismo de pesquisa para obter uma descrição da vulnerabilidade. A seguir, na porta 80, que segue a mesma metodologia da porta 21 descrita acima, mas agora referindo-se ao serviço Hypertext Transfer Protocol (HTTP) executado num serviço apache. Após o relatório NMAP, também no porto 443 foi detectado uma instalação do GlassFish, usando o serviço Secure Socket Layer (SSL), provavelmente porque os sites hospedados no Glassfish usam certificados SSL/TLS para proteger os dados trocados entre os servidor e o user. Além disso, a porta 8080 possui a mesma instalação do que o serviço na porta 443, mas hospeda a página Web que está sendo analisada. Também descobriu que os por-

tos 8081 e 8082 estão abertas, usando o serviço SSL que também hospeda as páginas Web que estão sendo usadas para o projeto EPOS. Relativamente à análise usando o ZAP, foi verificado outro aplicação Web, também dentro de EPOS, chamado *GNSS Products Portal* [12], com opções de força de ataque alto assim como o threshold

Esta aplicação recolhe dados pessoais devido ao seu mecanismo de autenticação, de modo que as perguntas do GDPR foram respondidas, dando o resultado final de 8 não estarem em conformidade de um total de 31 pontos. É importante realçar que essas perguntas foram respondidas por José Manteigueiro do projeto de EPOS, que é o desenvolvedor do sistema de autenticação. Para os pontos que o sistema não cumpre, foram apresentadas algumas sugestões, embora alguns pontos não as tenham, uma vez que as perguntas são explícitas o suficiente para saber o que mudar.

Também é relatado o uso de duas cookies quando um usuário acessa a aplicação Web, a *laravel_session* e o *XSRF-TOKEN*. Esta primeira é usada para identificar uma instância de sessão para um usuário e a segunda para impedir que eventos não autorizados sejam executados em nome de um usuário autenticado usando cookies de autenticação. Em seguida, foi detectado o uso de um gerenciador de cookies para gerenciar o consentimento e como o analisador de cookies desenvolvido tem em conta esta situação, foi realizado o clique para aceitar o consentimento e foi obtido o cookie *CookieConsent* com um stamp value. Juntamente com essas informações, vem as informações se as cookies tem o header de for *HttpOnly* ou não. Este campo é definido no cabeçalho da resposta e ajuda a reduzir o risco de client side scripting. Nesta situação, apenas a cookie *laravel_session* como o *HttpOnly* definido como true. Relativamente ao Wapiti, em ambos os testes feitos, não reportou nenhuma vulnerabilidade.

Concluindo, foi estudado o impacto que o regulamento GDPR pode ter sobre empresas com menos recursos, afetando sua capacidade de implementar e entender as mudanças necessárias para estar em conformidade. Isso também representa um problema devido ao aumento do uso de dados pessoais para desenvolver soluções personalizadas para cada user. Com isso, a busca por novos processos para lidar com grandes quantidades de dados e o uso de novas tecnologias baseadas em mineração e análise de dados também aumentaram. Por isso, o GDPR apareceu para proteger e dar mais direitos os titulares dos dados pertencentes à EU.

Durante todo o estudo, foram identificados vários problemas relacionados ao GDPR, por exemplo, a dificuldade de extrair do regular o significado de certos pontos. Dos artigos foi tirada a idéia de organizar o GDPR em sete princípios e foram escritos os pontos mais importantes associados a cada princípio. Além disso, e como é um dos princípios mais importantes, a integridade e a confidencialidade, foi organizado um conjunto de ferramentas responsáveis por encontrar falhas de segurança em aplicações web. O resultado final foi a PADRES que abrange os regulamentos GDPR resumidos em uma checklist junta com sugestões, um conjunto de duas ferramentas para a avaliação de segurança e também um scanner de cookies. No final, e com base na saída da ferramenta, um relatório detalhando esses resultados.

A primeira conclusão possível de ser feita é o facto de que todas os objectivos foram cumpridos. Desde os pontos de regulamentação extraídos às sugestões, incluindo um scanner de cookies e uma avaliação de segurança concisa, compondo todas as respostas para os objetivos definidos desde o início.

Embora os pontos de cada princípio e as sugestões correspondentes tenham sido extraídos por alguém sem experiência em direito, toda a documentação existente disponível era acessível e compreensível, mas, para fornecer essa ferramenta a qualquer empresa, esta parte deve ser revisada por alguém com conhecimento devido a suas complexas restrições e cláusulas ocultas. Alguns deles também requerem uma pesquisa posterior para complementar as informações já

obtidas.
Em relação aos resultados obtidos e analisados, é possível afirmar que a PADRES se comportou conforme o esperado. A análise GDPR no EPOS evidenciou que eles estão em conformidade com a maioria dos pontos, no entanto, a implementação das sugestões mencionadas também pode ajudá-los a obter mais conformidade, especialmente no princípio da explicação e demonstração, onde se espera mostrar o que foi feito para estar em conformidade. As ferramentas de avaliação de segurança, especificamente de ZAP, relataram algumas vulnerabilidades de falsos positivos, o que significa que além disso, há a necessidade de testar manualmente essas vulnerabilidades. Além desses. as vulnerabilidades restantes não eram críticas, sendo a maioria relacionada a parâmetros ausentes nos cabeçalhos das solicitações. A desvantagem das ferramentas de segurança foi o fato de o wapiti não reportar nada, mesmo quando os parâmetros foram alterados para tornar os ataques mais fortes. Mas como os resultados do ZAP foram falsos positivos, o relatório do Wapiti pode estar correto. O objetivo de usar duas ferramentas que basicamente fazem o mesmo era exatamente este para ter redundância. Além disso, o relatório NMAP pode fornecer uma visão geral da infraestrutura e das vulnerabilidades associadas aos serviços em execução nas portas encontradas.
Concluindo, a ferramenta pode ser útil ao fazer as primeiras abordagens GDPR, mas com base nos resultados, não é possível confiar inteiramente nela, exigindo mais informações de especialistas. O mesmo se aplica às avaliações de segurança. Com base no relatório, tomar as decisões necessárias para investigar as vulnerabilidades de forma mais exaustiva, se necessário. Além disso, a ferramenta deveria ter sido testada em outros aplicativos para obter mais resultados e, a partir daí, concluir quais são os princípios em que é observada uma maior ausência de conformidade e também ver quais são as vulnerabilidades mais comuns encontradas e combiná-las no OWASP top 10.
Com base nisso, para o futuro é importante ter esses dados para extrair mais sugestões e pontos de forma a obter um relatório mais expressivo. O ideal, seria um sistema capaz de extrair essas informações automaticamente, contudo nos artigos estudados, essas ferramentas já existem, mas não dão o resultado esperado [6]. Outra funcionalidade importante seria uma capaz de calcular o nível de anonimização de um determinado conjunto de dados com base no nível de entropia [14] ou com base nos algoritmos introduzidos aqui [15]. Além disso, o uso de mais ferramentas de avaliação de segurança, como a Arachni, pode introduzir mais redundância e possivelmente encontrar mais vulnerabilidades ou ajudar a evitar falsos positivos. Muita das ferramentas dispões de várias opções de configuração, por isso juntamente com a escolha das ferramentas para fazerem o scan seria também útil uma interface para permitir que o user escolha as definições dessas ferramentas.

Abstract

Since May 2018, companies have been required to comply with the General Data Protection Regulation (GDPR). This means that many companies had to change their methods of collecting and processing EU citizens' data. The compliance process can be very expensive, for example, more specialized human resources are needed, who need to study the regulations and then implement the changes in the IT applications and infrastructures. As a result, new measures and methods need to be developed and implemented, making this process expensive.
This project is part of the EPOS project. EPOS allows data on earth sciences from various research institutes in Europe to be shared and used. The data is stored in a database and in some file systems and in addition, there is *web services* for data mining and control. The EPOS project is a complex distributed system and therefore it is important to guarantee not only its security, but also that it is compatible with GDPR. The need to automate and facilitate this compliance and verification process was identified, in particular the need to develop a tool capable of analyzing applications *web*. This tool can provide companies in general an easier and faster way to check the degree of compliance with the GDPR in order to assess and implement any necessary changes.
With this, PADRES was developed that contains the main points of GDPR organized by principles in the form of *checklist* which are answered manually. When submitted, a security analysis is also performed based on NMAP and ZAP together with the cookie analyzer. Finally, a report is generated with the information obtained together with a set of suggestions based on the responses obtained from the checklist.
Applying this tool to EPOS, most of the points related to GDPR were answered as being in compliance although the rest of the suggestions were generated to help improve the level of compliance and also improve general data management. In the exploitation of vulnerabilities, some were found to be classified as high risk, but most were found to be classified as medium risk.

Keywords

GDPR, Vulnerabilitiess, Segurity, Compliance

Contents

List of Figures

List of Tables

Acronyms

AES Advanced Encryption Standard

API Application Programming Interface

CSRF Cross-site request forgery

CSV Comma-separated values

CVE Common Vulnerabilities and Exposures

CWE Common Weaknesses Enumaration

DDoS Distributed Denial-of-Service

DOM Document Object Model

DPD Data Protection Directive

DPIA Data Protection Impact Assessment

DPO Data Protection Officer

DPR Data Protection Principles

ECC Elliptic Curve Cryptography

EEA European Economic Area

ENISA European Union Agency for Network and Information Security

EPOS European Plate Observing System

EU Europe Union

FTP File Transfer Protocol

GDPR General Data Protection Regulation

GLASS Geodetic Linking Advanced Software System

GNSS Global Navigation Satellite System

GPS Global Position System

HTML HyperText Markup Language

HTTPS Hyper Text Transfer Protocol Secure

HTTP Hypertext Transfer Protocol

ICO Information Commissioner's Office

IoT Internet of Things

IP Internet Protocol

IT Information Technology

JSON JavaScript Object Notation

MAC Media Access Control

MITM Man In The Middle

MVC Model-View-Controller

MVVM Model-view-viewmodel

NMAP Network Mapper

NSE Nmap Scripting Engine

OECD Organisation for Economic Co-operation and Development

OSS Open Source software

OS Operating System

OWASP Open Web Application Security Project

PADRES PrivAcy, Data REgulation and Security

PbD Privacy by Design

PECR Privacy and Electronic Communications Regulations

PPDM Privacy Preserving Data Mining

PPDP Privacy Preserving Data Publishing

REST Representational State Transfer

RINEX Receiver Independent Exchange Format

RNG Random Number Generator

RSA Rivest-Shamir-Adleman

rxJS Reactive Extensions for JavaScript

SDK Software Development Kit

SFTP SSH File Transfer Protocol

SPA Single-page application

SQL Structued Query Language

SRS Software Requirements Specification

SSH Secure Shell

SSL Secure Socket Layer

TLS Transport Layer Security

URI Universal Resource Identifier

URL Uniform Resource Locator

USA United States of America

WASC Web Application Security Consortium

XML Extensible Markup Language

XSS Cross-site scripting

ZAP Zed Attack Proxy

Chapter 1

Introduction

1.1 Scope

In today's world, data can be extracted from practically everything, from clothing with small wearable devices to browser history, data is present everywhere. It has become a very useful and powerful asset to companies and that only means one thing, knowledge. The ability to target advertisements to specific people accordingly to their tastes, the intelligence behind the recommendations systems found on streaming services, like *Netflix* or *Spotify*, or the capacity to study irregular heart rhythms, found on Apple Watch, makes this asset one of the most valuable resources available for companies, allowing them to make a lot of money. It also allows the development of applications and tools that make our lives easier, in a way that we do not mind to give our personal information, to have services that are tailor-made to us. Since every little piece of technology share its data, become extremely important to make sure that, not only the communications are secure, but also the place where the information are stored needs to be virtually and physically secure.

This interaction between human and computer keeps increasing and has become so natural that it seems to be "invisible". This concept is defined by the term, *UbiComp*, which stands for Ubiquitous computing. Was introduced by Mark Weiser in his paper "The Computer for the 21st Century" [1]. Starts by declaring "The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it". This quote, express perfectly what is an *UbiComp* system. These systems, since they are directly connected with our way of living, they collect and process our data, some of it being private.

Since it seems a good trade-off between giving our data, often freely, in order to receive some personalized services, a problem emerges when privacy is violated. With this personal information that should not be available to strangers, can be accessed by anyone with good or bad intentions. In order to protect and regulate the privacy of EU citizens data, since May 2018 it has become obligatory to be compliant with the General Data Protection Regulation. The GDPR aims to give data owners the possibility to control and protect their own data. Companies that are not compliant with GDPR can face fines up to 20 million euros or 4% of their total worldwide annual turnover [23].

1.2 GDPR

GDPR was adopted in 2016, but before other guidelines and legislation existed such as the a directive called European Data Protection from 1995. This directive was the GDPR predecessor, and was created to be in conformation with the Article 8 of the Charter of Fundamental Rights of the European Union [2]. Its purpose was to create a framework that could guarantee the security and freedom of an individuals personal data across all the EU countries and also serve as a guideline on how data should be stored, processed and transmitted [3].

With the advances in technology and globalization, new challenges emerged regarding the data protection [24]. Those challenges required the development of a new approach which could guarantee the right of data protection for individuals and their personal data.
The concept of personal data is important to define because is one of the key concepts for the protection of individuals [24]. Personal data is ”any data that can be linked to a specific person” [17], linked directly or indirectly. The term ”any data” includes personal identifiers such as full name, national id number or indirect identifiers like Internet Protocol (IP) address or photos. If data does not have any of these identifiers then data is called anonymous [17] and in that case the GDPR does not apply [25]. Even though the data is anonymous, it might be re-identified using a technique called de-anonymization.

1.3 Security

The fields of study, GDPR and security, complement each other. No company can be GDPR compliant if its data is not secure. However the converse is not true. Data can be secure without being GDPR compliant however to be compliant data must be secure.
The section 3, will introduce the base concepts of what can be called personal data an what can be done to achieve its security. However, security is a broader area and can be implemented using different techniques and mechanism which will also be discussed in that section.
Yet, when security is not well implements or when software best practices are not used properly, the systems become vulnerable to be exploited. With this, OWASP provides a list with the most common vulnerabilities found in web application. Since we shall be focused on applying our approach to applications that consists not only of web sites and data portals but in general web applications, therefore it is important to test them against the most common vulnerabilities. Complex distributed web application also usually rely on a mixture of Databases and web server applications, proxies and gateways and thus it’s also important to test them, using auditing tools that are open-source.
OWASP is a not-for-profit group, which aimed in the beginning to raise awareness between developers and managers about the risks associated with web applications. Now has become an application security standard. Since one of his core fundamentals is that their techniques and materials are freely and easily accessible on their website [8], anyone can use it to improve their web application’s security.
One of their most known project is the OWASP Top 10. As the name suggest, it is a compilation of the ten’s most critical risks in web applications. This is possible due to data submissions gathered from companies specialized in security and then the items of the list are selected and ordered decreasingly, regarding the combination with ”consensus estimates of exploitability, detectability, and impact” [9]. This report is recommended to be incorporated in company’s security reports in order to minimize and mitigate security risks. [10].
Besides OWASP, it has become a current practice to use OSS applications running on top of the OS Linux. This direction that is being taken, can be justified, because OSS’s source code will be exposed to independent assessments making it more likely to have bug fixes and also a very important point is that is generally free. So, using OSS to look for vulnerabilities can be really important, because it will probably find them, while using proprietary software may not identify them intentionally.

1.4 EPOS

Solid earth science plays an important role in our society by studying diverse subjects such as geology, seismology, geodesy. This is done by enabling data about earth sciences from various research institutes across Europe to be shared and used, in order to be monitored so it can help us to have a better understanding of the dynamic and complex solid-Earth System. The European Plate Observing System, EPOS, "is long-term plan to facilitate integrated use of data, data products, and facilities from distributed research infrastructures for solid Earth science in Europe" [11]. In order to implement that integration, a distributed web software called GLASS, Geodetic Linking Advanced Software System, has been developed for that process.

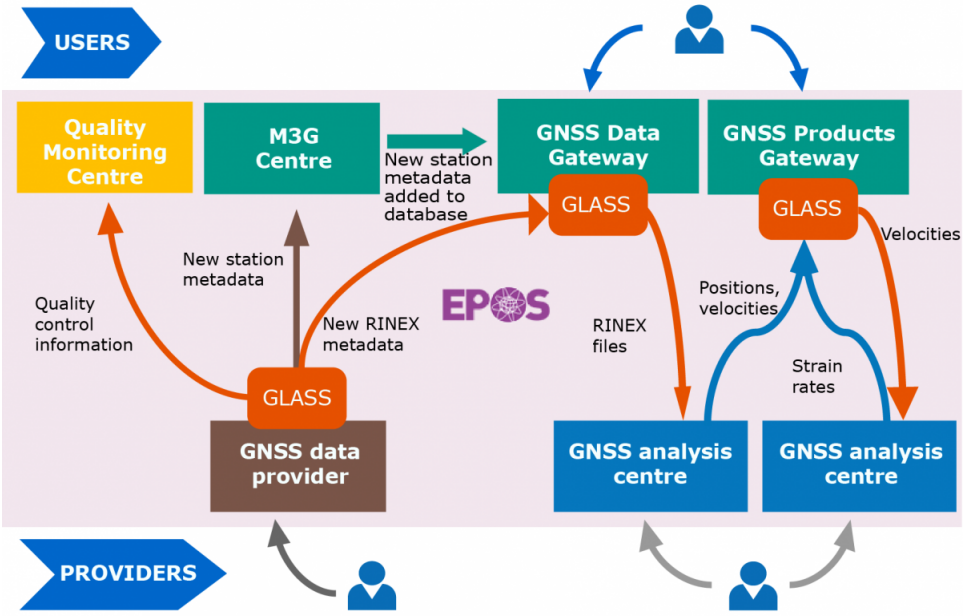


Figure 1.1: EPOS Diagram, summing its up the workflow [16]

Dissecting the figure above, data is supplied by the providers, which are agencies that collect raw data, and is introduced in the system through the GLASS software. This data is associated with a Global Navigation Satellite System (GNSS) station operating in Europe and is available in Receiver Independent Exchange Format (RINEX) file. Then, in order to validate it and become available in the system, the M3G [26] and Quality Monitoring Center After softwares are responsible for that analysis. If the data meets the standards and it is approved, is converted into products on the analysis centre. These comprise the positions, time series, velocity fields and strain rate fields. Then, those results can be monitored and seen in the portals by the users. Specifically in the GNSS Products Gateway and GNSS Data Gateway, which are going to be used as typical scenario and used to test the solution developed, that communicate with the GLASS through its API.

The EPOS GNSS platform also includes authentication mechanisms. Analyzing such a platform will will enable us to develop an approach for analysing GDPR with respect to such web applications. Also, having an overview of what type of data is gathered about the users while navigating through the data portals and also the data stored about the owners and managers of each Global Position System (GPS) station, that are used to collect the raw data about the Earth, will serve as a case study for the tool to be developed.

1.5 Goals

GDPR has become a top priority for organizations but also has become a problem. Some companies are not prepared for the changes that have to be made and are not aware of the consequences that such non compliance can bring to them. Studies have found that these problems happen due the fact that the actual regulation is ”vague, ambiguous and verbose”, meaning that anyone who does not have the proficiency required can find understanding the regulations very difficult. For example, the GDPR law says that companies must provide a reasonable level of protection of personal data [6], but the word ”reasonable” is not well defined. Also ”privacy by design” is promoted, without having a proper guide of how it can be achieved. The learning curve necessary to comply with the GDPR is generally time-consuming and cumbersome, making this a costly process especially for small and medium organizations that do not have a legal department or can not afford a legal advisory. As well as this, there are also two major problems that engineers and developers come across when trying to implement legal compliance [7]. The first is about determining which regulations can be applied and the second is related to the ability to develop the necessary policies that could enable compliance to those regulations, especially when extracting requirements from legal texts can be an error-prone job.
Since EPOS is a complex and distributed system it is important to assure not only its security but also that is GDPR compliant. This was the main motivation behind this work, namely the need to automate and facilitate this process.
The main objectives of this thesis are in particular to develop a tool and work-flow methodology that will be able to analyze complex distributed web applications not only for one specific case, namely EPOS, but for web applications in general. Such a tool offers the promise of an easier and faster way of checking the degree of GDPR compliance, in order to then implement the necessary changes.
To achieve this the tool named PADRES was developed. Using this tool an analysis of the web application needs to be done and then, accordingly to the result, a classification is given and a final report generated about what can be improved. The analysis can be seen as a survey that a system administrator or developer must answer manually. The specific questions contained in the survey have been created during this work. These questions are constructed from an analysis of the GDPR made during this work. As there is no such thing as privacy without security the tool that was developed also includes a search for vulnerabilities. This is made by integrating a combination of open source tools, like NMAP or ZAP, in order to test the platform against the known risks mentioned in the OWASP Top 10 Application Security Risks of 2017 [27]. The result of this security analysis will also have an impact on the final classification. In the end, a report will be generated with everything that needs to be improved or changed.

1.6 Document Organization

The rest of this document is structured in the following way:

1. The first chapter – **Introduction** – introduces the key elements that will be developed in others chapters as well as the overall scope and goals of this work.
2. The second chapter – **Privacy and Security : A Theoretical Foundation** – since this is a key topic in GDPR and also a major subject in today’s applications,it is important to understand what is being done to prevent security breaches and the also to study the most common vulnerabilities. This section also discusses methods to increase the data subject’s privacy.

3. The third chapter – **GDPR: History and State of the Art** – gives a background about data privacy and how the GDPR emerged. In this chapter we analyse related work and articles concerning GDPR.

4. The fourth chapter – **Application Development** – introduces a case study that motivated this work. Methodologies used to stored and collect personal data are discussed. This chapter also presents our approach to the resolution of the problem and discusses the architecture and implementation of the tool developed.

5. The fifth chapter – **Tests and Results** – discusses the application and results from two case studies where our approach and tool were applied.

6. The sixth chapter – **Conclusion and future work** – will discuss what was achieved and also what were the difficulties while building the application and studying the literature. Finally the improvements that can be done in the future are listed together with other future work.

Chapter 2

GDPR: History and State of the Art

2.1 Introduction

In this chapter, a background about data protection will be given, specifying some key topics regarding this subject, in order to understand why and how GDPR emerged. So, brief overview of GDPR was already done in the section GDPR, therefore this chapter will be more focused on giving an in depth analysis of GDPR and also an approach on how to solve this emerging problem through security and therefore privacy.

2.2 Chronological Background

The fact that data protection has become a significant issue, concerning the privacy of people, is not a recent phenomenon. Indeed, in 1890 an article entitled ”Right to Privacy” [28] was published. The writers admit that, from time to time it is necessary to change the law, in order to satisfy the requirements of the communities. By that time, it was already identified that ”recent inventions and business methods” were happening and becoming more sophisticated and the need to protect the individuals from public exposure was emerging. With that, a new term appeared, ”to be left alone”.
That was just the beginning of what came next. In 1948, with the adoption of ”The Universal Declaration of Human Rights” [29] as a basis for all the people from all nations, the article 12 entitled ”Right to Privacy, was included, defining that ”no one shall be subjected to arbitrary interference with his privacy”. The key part on this citation is ”arbitrary interference”, which can be understood, and following the justification from here [30], as a rupture in the law, such as, give access to personal information or in the data collection, which led to for example, financial loss, humiliation or loss of dignity. The justification used above, also presumes that, the collection and storage of personal data must be regulated and therefore measures must be taken to ensure and audit this requirement.
The next stop in this time line dates is 1995. In that year, the European DPD. was created. This directive was adopted in order to protect the EU citizens data. The structure of this directive, was established based on the eight principles published in the document ”Guidelines on the Protection of Privacy and Transborder Flows of Personal Data” published by the Organisation for Economic Co-operation and Development (OECD). Such guidelines, gave a general recommendation to governments and business, on how to collect and manage personal information [31]. Some of those principles were already explored and examined on the section GDPR, since the GDPR is an improvement of this directive. The major changes which are worth a mention, are shown on table 2.1. Here is important to notice the difference between a directive and a regulation, in this case, the DPD and GDPR respectively. A directive is a legislative act [32], where is up to the countries of the EU, to decide how to achieve the goal, which was given by the EU. The regulation on the other hand, must be implemented in full throughout across all the EU.

Between 1995 and 2016, when GDPR was discussed and approved by the EU parliament, were adopted other directives such as the Directive on Privacy and Electronic Communications [33], concerning the processing of personal data. This directive, also known as *e-privacy Directive*, sought to complement the directive from 1995 [34] and the main point was how electronic communications should be addressed regarding privacy rights. More specifically, it covers fields of marketing calls, texts, and emails, which are known as digital marketing. The Privacy and Electronic Communications Regulations (PECR) has restricted such unsolicited marketing, ensuring that this is only feasible if approval is provided. This consent, must be provided freely and is suggested to be done by a ticking box in a website, as shown in point 17 [35]. Currently, this directive works alongside the GDPR. This implies that companies must be able to comply with both, as they share some of the same principles, more specifically the consent one. Such principle, under the GDPR, is only applied if the processing of data is related to personal information. The PECR instead, applies even if data is not personal. It was also during this time that some newsworthy data breaches and leaks occurred, some more critical than others. One of the most famous and that caused a lot of impact, was the *Yahoo* one, where data like, personal name, date of birth and email address, were made public. It was estimated that more than 1.5 billion accounts were leaked, making this the biggest leak ever recorded [36]. It was also found, in November 2018, that nearly 500 million people's personal data gathered by the *Marriott* hotel chain were compromised. Now, almost one year after that event, the Information Commissioner's Office (ICO) aims to fine them more than 99 million pounds, under the GDPR law [37]. Even though the *Marriott's* data processing is located outside of the EU and some of the data leaked is related to EU citizens, the GDPR is also applied, as explained on section GDPR. These types of leaks can lead to identity theft, and can then be used to commit crimes such as "application fraud", that consists of applying for a credit with only a name and a social security number, or to obtain credit card numbers [38], which is called "account takeover". This, is not only a problem for the people who have been stripped of their identity, but also contributes to the loss of confidence in these corporations. Finally, to protect the EU citizens, in 2018 the GDPR was adopted, replacing the DPD 95/46/EC changing some concepts and introducing new ones [39], as can be seen on the following table.

The table 2.1 contrasts and compares the two legislative acts. In the first column the group of rules where changes were made is shown. The second one describes the key points of DPD regarding that group. In the last column we have the new or the updated rules. With this table it is possible to have an overview of what changed.

By taking a closer look at the table 2.1 is possible to see that the group personal data, which was firstly defined as described, now as wider definition, containing not only the previous points, but also the ones mentioned, as well as, the points such as economic status, cultural identity or online identifiers, like cookies or IP addresses of the data subject. This change, in order to protect the EU citizens data, makes a big difference for companies that for example are using profiling methods, so they can do targeted marketing as now they have to ask for consent.

Going forward to the next row, we have the subject Individual rights. Both legislative acts, share the same point, the consent, and there are no considerable differences in the definition of them. The following points under the GDPR column, add more rights beyond the ones provided by the DPD. This subject, has an important role under this legislative act, since it give us, the individuals, the ability to protect ourselves and to be alert to possible ways to contravene the law, so it will be given a special attention in chapter 2.3, in order to understand all the rights and the means to exercise them.

Table 2.1: Comparative table between DPD 95/46/EC and GDPR

Changing points	**DPD**	**GDPR**
Personal data	• Smaller criteria with points as name, photo, email address, phone number and any personal identification number	• Broader definition adding points such as IP address, mobile device IDs, geolocation and biometric data
Individual Rights	• Consent • Right of access • Right to object • Right of rectification	• Consent • Right to portability • Right to restrict processing • Right to be forgotten
Data Breach Notification	• Different data breach notification laws for each member states	• Obligation to notify the supervisory authority
Security	• Level of security appropriate to the risks • Technical security measures	• Technical and organisational measures to ensure a level of security appropriate • Pseudonymisation and encryption • Privacy by Design • Designate a DPO

Moving to the second to last point on the table 2.1, in the DPD, EU member states are free to adopt and apply their own Data Breach law. This meant that in the eventuality of a data breach companies would have to do research for each EU member state law, in order to be in compliance with them thus increasing the complexity, time and effort needed by these companies that have suffered a data breach. With the advent of the GDPR, more precisely Article 33, there is only one law for all the members of EU, which states that the controller is obligated to inform the supervisor authority, no later than 72 hours, after becoming aware of the data breach. This is not obligatory if the ”data breach is unlikely to result in a risk to the rights and freedoms of natural persons” [40]. If the national authority is not informed within 72 hours, the notification must be followed with the reason of the delay. This notification must contain at least the possible implications, the name and contact details of the data protection officer or other contact point who may have more information and also the approximate number of data subjects affected. Besides this the action taken to mitigate this data breach must be documented and the supervisor authority must be able to verify the compliance with this article.

The last considerable upgrade to DPD is in the security point, as shown in the last row of the table 2.1. Even though it is one of the main subjects in computer science with diverse subfields,

when is analysed in the sense of GDPR, the topics approached are the security of personal data and data privacy.

The Security of personal data has an important role under the GDPR, as it had on the DPD, with the differences now being that the controllers are obligated to implement measures, which are detailed in a few dedicated sections, more specifically sections two and three [40]. These sections introduces obligations and recommendations for the data controller and processor, such as the implementation of technical measures ”to ensure a level of security appropriate to the risk” [40] or the need to have a Data Protection Impact Assessment (DPIA). Those measures, contain technical methods like privacy by design, encryption or pseudonymization and were already studied on subsections under the section 3.2. It is also introduced the obligation to appoint a DPO. Some companies are not obligated to designate this position for now, but it recommendable to name a member internally to help on the GDPR implementation [41]. His main tasks are, as stated on article 39, ”to inform and advise the controller or the processor and the employees who carry out processing of their obligations pursuant to this Regulation and to other Union or Member State data protection provisions” [40], to ”give advice and recommendations to the institution about the interpretation or application of the data protection rules” [40] or to ”handle queries or complaints on request by the institution, the controller, other person(s), or on her own initiative” [40]. The one appointed to this role, must have knowledge in data protection, as well as, a good understanding of the way the company works. It is also worth to point that, must be no conflict of interests between the role of DPO and other tasks performed by him. In order to prevent that, the DPO should not report to a direct superior and have the responsibility to manage their own budget [42].

When studying about GDPR is important to be aware of some definition and concepts introduced. In the introduction section 1.2, was introduced the personal data conecept, but is also important to dissect the roles of each actor inside GDPR. Currently there are three defined and the relations between them can be seen in the picture 2.1.

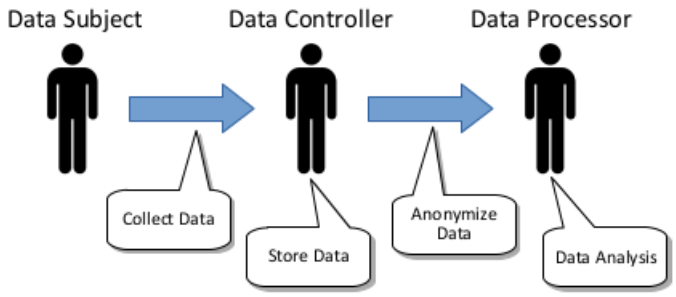


Figure 2.1: GDPR Roles [17]

The data subject is the person whose personal data is being collected or processed. Since the GDPR aims to protects those individuals, they have rights that goes from GDPR Article 12 to Article 23 [4]. After data being collected, the data controller takes the responsibility, which accordingly to Article 4 of EU GDPR, means that he is the person which determines the purposes of data processing. Finally, this processing, is done by the data processor on the behalf of the controller. Giving an example, a supermarket is the data controller, so it collects the data of its clients when they go buy some groceries, then another organization, having the role of processor, stores and process the data given by the controller. Both the controller and processor are responsible for handling the personal data.

Table 2.2: GDPR principles [6]

1	Lawfulness, Fairness and Transparency
2	Purpose Limitation
3	Data Minimization
4	Accuracy
5	Storage Limitation
6	Integrity and Confidentiality
7	Accountability

For data processing, GDPR defines a set of seven principles [5], which can be seen in the table 2.2, being only the most relevant described bellow.

Before starting collect data, a consent, as described in the Article 6, from the data subject must be given, in order to process his or her data. So, the data controller must be able to demonstrate such consent. Also, the data subject can withdraw at any time his consent and this process must be as easy as to give consent. Regarding to this, there are some exceptions that apply. When the processing is necessary for compliance with a legal obligation, when processing is done in order to protect the vital interest of the data subject and by this, its understood only interest that are essential for someone´s life an only applies to matters of life and death [43], for example in case of a emergency medical care and the data subject is not capable of giving consent. Other exception occurs when the processing is necessary for public interest or for an exercise of official authorities like reporting crimes, taxation or public health.

The last exception that is worth to mention, is processing for the purpose of the legitimate interests pursued by the data controller or a third party. GDPR does not give a definition of legitimate interests so it can have an adjustable interpretation in order to be applied in different situations. An example is when a company needs to process data in order to assure the security of the network and the security of the information.

Also when giving consent, is data controller's responsibility to specify the purpose of the processing. This, must be limited, clear and well documented so any individual can have and understand it. Also is important to mention that if consent was firstly given for some purpose and now the data processor needs that data for other purpose, the data controller has to check if there are compatibilities between them. If there are not, a new consent must be given.

Following this last topic, a consent must be given freely, so there can not be any element of pressure or compulsion. The following quotation from the Article 29 Data Protection Working Party [44], emphasizes that: "The element "free" implies real choice and control for data subjects. As a general rule, the GDPR prescribes that if the data subject has no real choice, feels compelled to consent or will endure negative consequences if they do not consent, then consent will not be valid. If consent is bundled up as a non-negotiable part of terms and conditions it is presumed not to have been freely given. Accordingly, consent will not be considered to be free if the data subject is unable to refuse or withdraw his or her consent without detriment".

To end this principle of consent, there are some special categories such as race, health or genetic data which require explicit consent. This can be achieved by an express statement of consent given by the data subject or even further, that statement needs to be signed by the data subject, so it can be used as an evidence in the future.

Moving to the next principle, data minimization, it aims to limit the collection, storage and usage of personal data, in a way that its only relevant, adequate to achieve the purpose for which was collected. This means "nothing more than what is indeed needed" [45]. To accomplish such thing, some questions can be specified like "Is there a way of achieving this purpose without

having to collect the data?" or "How long will I need the data for to achieve the purpose?" [46]. By answering this questions, besides being GDPR compliant, it also helps to reduce the storage needed, because data, as the time passes, becomes less useful. On top of that, it also protects against possible data breaches, for example in South Korea, data minimization policy is being used in practice to avoid the repetition of what append in past [2]. The term data breach is not only about losing control of data. GDPR defines it as "breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed" [47].

Also is worth to mention the technical measure*pseudonymization*, used to accomplish data minimization. This method can be seen as an advanced form of encryption that does not require passwords or encryption keys [2], consisting in replace personal identifiers with randomly generated identifiers. This process, should be firstly initialized by data controller, before passing the data to the data processor. When the data controller does this step of replacement, a master table, also know as *lookup* table, is generated, being in the future possible to turn the random identifiers to the original ones.

Using this process, companies which are processing that data, can do it more efficiently, simple because there are no encrypted files and most importantly the risk of processing personal data is reduced. Another good point on having *pseudonymization* or encryption implemented, is that companies does not have to report their data breaches to the affected individuals, which accordingly to the Article 34 of GDPR needs to be done if a data breach may result in "in a high risk to the rights and freedoms of natural persons".

Even though sometimes is not possible to anonymize data, if during the processing of that data, a data breach occurs and threatens the rights and freedoms of persons in order to be in compliance with GDPR, it might be required to have a data protection impact assessment (DPIA). The purpose of this assessment is to help identify and to minimize a project's data protection risks. With this, a new principle, which is not more than extension to the last one mentioned, called storage limitation emerges, saying that data only should be stored only for the duration period which is necessary.

Related to data storage, is the problem of sending data to third-countries, which is defined as all countries that are not EU's members and do not belong to the European Economic Area (EEA) and GDPR forbids the sending. Among those third-countries, the European Commission has defined a group of countries which are considered secure including Switzerland and the United States of America (USA) which means that data transfer to those countries is permitted [48]. For a country to be considered secure, it has to provide a level of data protection similar to the EU law. If a country is considered unsafe, it does not mean that the data transfer can not be done. It belongs to the data controller the responsibility to ensure that the data is sufficiently covered by the recipient. This can be achieved using standard contractual clauses [48].

With this, the Article 37 of GDPR, suggest the designation of a DPO by the data controllers and the data processors and must be an expert with knowledge "of data protection law and practices". Besides this is must be able to fulfill, at least, the tasks mentioned on the Article 39 of GDPR, like for example, "inform and advise the controller or the processor and the employees who carry out processing of their obligations" to comply with GDPR and "to other Union or Member State data protection provisions", or to "the assignment of responsibilities, awareness-raising and training of staff involved in processing operations, and the related audits" . Since it is not obligatory, this appointment must be done when the data processing is carried by a public authority with the exception of courts. Also is required when companies are constantly monitoring the data subjects on a large scale as part of their main activities or when the core

activities of the company consists about processing special categories of data or data related to criminal activities. If a company does not fit in one of the points mentioned before, it can volunteer in order to nominate a DPO.
The sixth principle aims to ensure that companies must use the appropriate security measures in order to protect the personal data, so it is often called the security principle and it is explained on Article 32. Those measures will be approach on chapter 3, where a deeper study will be performed.
Last, but not least, the accountability principle, which obligate companies to not only perform appropriate technical and organisational measures regarding data protection, but also to be able to demonstrate compliance with all the principles. This type of governance, requires the companies to create a cultural and organisational change for GDPR compliance [49]. It can be achieved by ”developing internal guidelines for employees” [49], as well as, provide training for everyone which will be performing data processing.

2.3 GDPR Individual rights

This topic is often divided into eight rights [50]. Some were mentioned in the GDPR column in table 2.1 just to give a glance. In this section will be given an explanation of how they can be used by the data subject and refused by the data controller, to all of them.

- We start with ”the right to be informed”. As the name suggests, it means the right to know the identity and the contact details of the data controller, to know the purposes of processing for which data was collected or to know the period for which the personal data will be stored, and if that is not possible, the criteria used to determine the period

  [40].

- Moving to ”the right of access”. This was firstly published under the DPD being defined as the right to ”every data subject the right to obtain from the controller” [51] the following information, without excessive delay or expense, the confirmation if their data is being processed, the data which is being analyzed and any information available regarding its origin, as well as the knowledge of the logic involved in automatic data processing. This right under the GDPR, shares the same points as the DPD, with the addition of the right to ”lodge a complaint with a supervisory authority” [40] or the right to know the implications of not providing the data if the data subject is obliged to do so.

- ”The right of rectification”. This brought to data subjects the entitlement to, without undue delay, the rectification of inaccurate personal data, as well as the right to have incomplete personal data [40].

- ”The right of erasure”, or as is often known, the right to be forgotten, explaining that the data subject can request their data to be deleted and the controller must have to obligation to do it. This request shall have a cause, that can be for example the unlawful processing of data, if the data is no longer required or if the data subject withdraws his consent. Moreover, ”taking account of available technology and the cost of implementation”, the data controller, ”shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those

personal data" [40]. Those grounds can not be applied if the processing of data is necessary for exercising the right of freedom of expression and information, for reasons of public interest in the area of public health or for the establishment, exercise or defence of legal claims [40]. This last sentence means that this right is not absolute and can be refused.

- Related to erasure and deletion there is the "right to restrict processing". Here the data subject can request their processing of data to be limited if one of the rules described next, may be applied. Those grounds are, the data accuracy is contested by the data subject, or if "the processing is unlawful and the data subject opposes the erasure of the personal data and requests the restriction of their use instead" [40], or if the controller does not need the data anymore, but "they are required by the data subject for the establishment, exercise or defence of legal claims" [40].Before the restriction is removed, the data controller must inform the data subject.

- Also created was "the right to data portability". This allows the data subject to request his data from the data controller, who must make it available in a "structured, commonly used and machine-readable format" [40], and has the right to give it to another data controller. Also the transmission of data to other data controller can be requested to be done automatically if technically achievable. This right can not be applied if the processing its subject of public interest or if it affects the rights and freedoms of others.

- The second to last right that will be focused on is "the right to object". This gives the data subject the ability to, at any time, object the processing of their data, if the processing has the purpose of direct marketing, which includes profiling. The objection can be refused by the data controller if it is proven that the processing has the compelling legitimate grounds which overrides " the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims" [40].

- Finally, the rights related to automated individual decision-making, including profiling, where the data subject shall have the right "not to be subject to a decision based solely on automated processing" [40]. This right can not be used if the processing is necessary for entering into or to perform a contract between the data subject and a data controller, also if there is an explicit consent. In both cases the data controller must implement measures to safeguard the data subject's rights and freedoms and legitimate interests.

2.4 GDPR related articles

The first paper regarding the topic GDPR, is entitled "The Grace Period Has Ended: An Approach to Operationalize GDPR Requirements" [6].
It was chosen, not only because that, by reading the abstract, it can be the extracted the quote, "it is difficult for practitioners to extract and operationalize legal requirements from the GDPR", which basically summarises the problem that was found and firstly referred on the section 1.5, but also that the goal aims to help the organizations to understand their obligations under the GDPR, by proposing a solution called *GuideMe* and being to proof able at the end, that their approach meets the recommendations of privacy experts. Whit this, it is expected that our approach to the problem will be stronger and well grounded.

Starting on the section one, an introduction is given. This, starts by analyzing some surveys in order to justify why the GDPR will be a problem to some companies, due the fact that, some of them are not aware of the implications that such regulation will impact their business. The reasons are divided into, some directors are not informed about it or being in doubt about their companies be covered by it or even more precarious, the inadequacy that such regulations have in terms of communication.

Besides that, the bigger companies will have less problems to be in compliance with GDPR, because they have more resources, both to understand and extract from the GDPR what needs to be changed and to implement them. On the other side, the small companies, having less access to law and technical experts, will have more difficulty to be in compliance, due to this task be demanding in financial and human resources [41], leading to potential fines in the end.

Moving to the section "Related Work", it is evaluated a various number of checklists tools, as well as decision support system to help the identification of legal requirements.

It is concluded that, those checklist tools, help to identify not only the discrepancies in the compliance of certain companies, but also suggest other measures to protect personal data. The missing point here is the lack of specif suggestions to be incorporated in enforcement of software systems. It is also the shortage of expert knowledge to help in the identification of specific data processing exercises or in specific scenarios.

In the paper is also highlighted a progress that has been made, in order to try to extract the legal requirements from the regulation, even though is was identified some problems such as the scenario context.

Regarding the data protection measures, is was also identified some studies and measures, such as, the pseudonymization to reduce the data subject privacy risks or the privacy by design approach with the implementation of privacy patterns. Despite those methods and approaches provide a good ground regarding the privacy of data, is noticed that them do not give information about the degree of compliance.

In order to combat the problems recognized before, the authors, in section three, suggest an six step approach to help to obtain detailed information on what to do, to be in compliance with GDPR.

In this section, the authors begin by interpret a case study, an university, which process personal data about students and staff. To have an overview of the system, they built an use case diagram, where it is possible to observe the the relationships between the use cases, actors and systems. This type of diagram, specifically on this case, allows to see what data is going to be collected, where will be used and what type of authentication will be used, which in this case is by finger printing, that is included in the biometrics type of data, so it has an special way of being processed and collected under the GDPR.

On the Assumption subsection, they begin by stating that "of the most important provisions of the GDPR in relation to data protection is Privacy by Design and by Default" and that under that methodology the most important obligation that can be applied are the Data Protection Principles (DPR)s, which were used to structure our approach to the GDPR given on subsection 1.2.

It is also important to highlight their own perception of those DPRs, which were seen as business requirements in order discard any possible subjective interpretation. This achievement is possible because they linked each DPR to various business requirements specified in a Software Requirements Specification (SRS) document.

Those type of documents, often written under the ISO/IEC/IEEE 29148:2011 [52] which aims to define how to build a good requirement by providing the attributes and characteristics that a

requirement should have, ”describes all the externally observable behaviors and characteristics expected of a software system” [53] in order to build a software that solves the user needs. One of the problems identified in the paper [52], is that it will be much more expensive to solve an error detected in the advanced phase of the development that it would be if that error was identified on the requirements phase. So, the importance and need to have a good SRS is really high.

The authors justify their choose because it help to decompose the DPR into more precise functionalities and also by giving an example of one of those requirements. These requirements used together with a glossary, which defines essential GDPR terms, have an important role in the development phase, because it will help the Information Technology (IT) professionals to have a better understanding on how to approach the problems. The example is grounded by the book [54], specifically on chapter 2.3.

Also Martin et al. [55] address this problem on their article by stating the lack of skills regarding data protection in software engineers, mainly because their usual skill set rely on working with data flow and database structures. Also, the engineers find difficult to translate regulation point into operational working items. In respect of GDPR it is explained that they often are unsure on what technical measures to use to meet the user’s rights or if portability means that they have to ”slit opening databases”. In order to answer these challenges, Martin et al. suggest the use of data protection principles, such as Privacy by Design (PbD), to meet the compliance necessary.

V. Ayala-Rivera et al. [6] approach is based on a 6 step guide and in each one it is explained what should be done by ever participant involved in it, starting by performing a data audit, in order to understand what type of data an entity holds, how the data is being processed and stored. After that step is done a gap analysis, responsible to understand the what needs to be improved by introducing new requirements. The next step is the planning and preparation and, based on the previous step, it is determined the privacy controls needed to fill the legal obligation. Here is performed the map between the legal obligations and the SRS, making easier for developers to interpret the changes. The following steps are the plan review, responsible to review the previous steps, the execution step, where the solution requirements are implemented. Finally, the evaluation step, responsible to ensure that all the solution requirements are satisfied.

The next article which is worth to create a good ground, is entitled ”EU General Data Protection Regulation: Changes and implications for personal data collecting companies” [41].

This article has as its main goal the study of the differences and upgrades that were done from the DPD to the GDPR and then to identify the changes that companies, which work with personal data, have to do, in order to be in compliance with GDPR. Those changes, will enforce companies to implement new methodologies and practices, which will represent an investment in legal experts, which, as identified before, can be a problem to ”small-and medium-sized organizations” [6], that may not have the resources to apply those changes.

This paper was also chosen due the fact that, the authors have identified and classified the implications of those changes and with that, they developed a framework addressing those adjustments and how to be prepared for the requirements, by examining the best practices and technical measures.

In the beginning of the the paper, the authors begin by mention that the the GDPR will affect all companies handling data from EU citizens, meaning that, even companies outside of the EU will also be obligated to comply with this legislation. Overall, with was built to meet the new challenges which are emerging, regarding personal data protection and online privacy rights.

After that, as well as on other papers, such the one analyzed before, it is acknowledged the

fact that, the measures that are needed and mandatory, to seek a better data privacy can be challenging. They give by example the privacy by design concept, as one of those measures, because it is a ”proactive approach”, which will help organizations the deal it privacy since the design of software systems. They summarize these measures, by stating that, moving towards the implementation of data privacy measures, organizations will need considerable amounts of time, resources and guidance, to do it.

The next paragraph, also has important information to highlight. The one, which was already identified before, is the ”lack of awareness and understanding” of the new obligations that GDPR demand. Since these obligations can be translated into practical measures, it will bring new challenges to companies, which for some of them, can be hard the achieve.

This paper, has the particularity to be more focused into companies that their business rely heavily into ”extensively collect and process personal data”, including data from financial services or health care, which are very sensitive types of data, so they need to take into account the GDPR requirements.

After the authors have introduced their methodology, where was specified two questions, in order to fully understand the problem and how to approach the changes introduced by the GDPR, by giving a deeper and enlightening study about all the GDPR articles, they created a new section called ”Practical implications of the GDPR”. This one, aims to identify the practical implications, which means what companies need to change or to start doing to fulfill the GDPR obligations. They define has starting point, the acquisition of knowledge regarding GDPR. Then, the training that companies need to provide to who will execute data processing tasks, as well as, the assignment of new responsibilities which may lead to the hire of new people with the know-how in this area.

To summarize those practical implications, the authors created a table, containing twelve key implications. Each one is composed by the practical implication and a succinct definition of the requirements needed to fulfill this implication.

For example, they identified that, one of those practical implications, is the need of ”consent on personal data usage”. The implementation requirements to accomplish it, are the need to ask for consent, the ability to show that consent by the data controller when requested, what type of information the consent must have and the need to implement a functionality to provide the data subject the ability to withdraw his consent.

With this, companies can have an overview of the changes that they need to do, as well as, a description of the requirements, making it easier, for some companies to address their compliment regarding GDPR. In this case, the definition of the requirements, regarding technical measures, can be difficult to interpret by IT developers. So, the solution introduced in the paper [6], where the authors linked each requirement to a business requirements using a SRS, makes it easier for developers to analyze. A combination of these two solutions, would provide a strong ground for companies which may have more difficulties to be in compliance.

The next article introduces how the GDPR will provide protection for privacy and also control the identity of users, in a specific scenario, more precisely, in Internet of Things (IoT). This paper covers the importance for companies to do profiling, but also the challenges that can result from that technique regarding the users privacy and protection on identity.

So, the paper entitled *Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR* [56], may give a different and critical view, over this news technologies that are emerging and also how the GDPR can help or difficult with the implementation of privacy and security in those systems.

The author starts by stating that the growing of this technology will keep to attract a lot of

investment and since it can be found everywhere, from smart homes to the healthcare industry, all that personal data collected from the devices that are used everyday, are required in order for the services associated with that devices to run properly.
This need to have tailored services to each user, leads that devices and services must be connected in order to exchange information, and this can bring risk to the users privacy. It is identified in the introduction of this article, that some of those risks can be realized by the users as invasive or unwelcome and may lead to discriminatory treatment.
The author identify a problem regarding the incapable communication of the risks associated with data processing that should be present in privacy policies. So, the author affirm that the GDPR might improve this situation by enforcing new data protection standards as well as principles for data processing.
The section two of this article, explains how to identification technologies are a central component inside IoT. Since one of the principles associated with IoT is the communication a data sharing between the all the connected nodes, the consistent identification of devices and users is required, in order to ensure that the communication is done between the intended target. With that, emerges the need to have an identity management, which are the technologies responsible to assign and manage the identities in order to establish trustful communications between IoT actors. These technologies can be divided into centralized and distributed, being the first one, the most used, with the justification of being better for users, but facing challenges such as scalability and cross-border governance.
In this paper it is also highlighted the use of biometric data to authenticate the users, being it done using iris recognition or electrocardiogram signals. Even thought they are considered probably more secure than the classic username and password scheme, they bring in new privacy risks.
The third chapter, enhance the ”tension between user’s information privacy (or control of personal data) and Iot”. The need of having the capability of profiling, in order to create a individual identity so it can have a personalized treatment inside the IoT, also brings the need of having devices constantly monitoring and collecting the users data. In order to achieve this personalization, data must be shared across the IoT controllers and also to third parties, so it is possible to link data from various types of sources, which is done most of the times without the consent of the users.
To protect the users, is stated on the paper, on chapter four, the introduction of GDPR, which will establish new standards for the IoT field, as well as, for the data controllers.
This chapter also identifies the privacy risks associated with IoT, being it the lack of control that users have over their own data, the insufficient knowledge about free given consent, the automated decision making and the combination of datasets. With the risks identified, the author is able to structured them into four groups of challenges, being the first the problem associated with invasive profiling, inferences and discrimination when the IoT controllers are able to link the users identities. The example that the author gives is very explanatory and alarming, because of the unawareness of the users and because it can be unfair to the users. The example is based on the data gathered by the *FitBit* device, that collects the user’s health data and also the user’s movement. With the combination of this two datasets can lead into privacy invasive inferences. They also give the example of third parties, such as employers, which are often interested in not only, but also private data, in order to have a better understanding of their habits. With the introduction of GDPR, where the author states the most important regarding this group, where it aims to give to the data subject more control over its own data.
The second group, lays on the use of personal data by third parties, that was shared from

a previous IoT device, leading the data subject to lost control over its own data. The data sharing, is often not consented by the device's owners, so they act on the user's behalf, sharing sensitive data defying the user's privacy, because it generate private data, like the location or the habits of the user. The third group, highlights the generation of information about the user, that could not have predicted when giving consent or setting access policies. Also in this group is mention the use of a privacy coach. A tool based on the comparison between the users privacy preferences and the downloaded privacy policy from an object, which as scanned and in the end giving a recommendation if the object meets the user privacy. The last group, focus on the limitation of the user to be able to do its own management of identity and profiling, leading possibly into breaches of privacy and to the lost of confidence between the users and the IoT providers. This group introduces the trust, between not only, the users and the devices, but also between devices and devices. The trust between the devices, is translated into the the authentication before doing the communication, which in the end will increase the user trust. The trust between the user and the device works in the opposite way. The author points that trust is often a prerequisite for systems and can be achieved by the demonstration of techniques to mitigate the risks. Also the need to have tools each are transparent on how it is possible to see all the interactions between their data, can increase the user's trust. The use of DPIA is now a requirement under the GDPR, because it helps to evaluate the risks of data processing and also proposes the plans to mitigate the risks.

The author accomplishes that, due the huge amount of data collected by IoT devices, their are a lot of risks associated with the profiling that is needed in order to provide to the final user, a better service experience. It the help of GDPR, it is now possible to lighten those risks, by giving more capacity to the data subject and also be requiring the implementation of techniques regarding the user privacy and security.

This article provides an excellent review of the most concerning risks associated with IoT and by linking them with the exact GDPR articles responsible to answer and mitigate them, companies can find here a powerful basis to deal with them, to by the end, be in compliance with GDPR.

Also in the area of IoT comes the next article [57] entitled "Enhancing User Privacy in IoT: Integration of GDPR and Blockchain". Even Though is again related with IoT, this article approaches a new method of data privacy, in order to give to the data subjects more control and information about their data and devices. In this case the authors used blockchain technology to achieve more transparency of privacy.

This article begins by stating the increasing number of wearable devices, such for military use and also for civilian usage. The challenge that comes with it is that those devices are collecting personal data and being surrounded by other ones, it is expected that they can communicate with each other, possibly requesting the personal data and sending it to third parties. The authors affirm that the use of GDPR can improve the user privacy in IoT.

The use of blockchain in IoT, is justified, in this article, with a reference to several articles with practical applications to some areas such as healthcare or communications, which in the end improve the "transparency, trust, and privacy". After that is stated that such approaches are missing the use of GDPR, since the design phase or to automatically verify the compliance when data is being processed.

On chapter 2 the authors, explain behaviour of an e-health monitoring system, where a patient is monitored by a vital device and also has body sensor which measures the blood glucose. This information is then sent to his smart phone, where the patient can see is general health conditions and also track the diabetes. If a read value is above the threshold, the smartphone connects to the nearest emergency centre, which can call for assistance. Besides that, every

read is kept on a local storage, in order to be accessed by a physician
Given that information, the authors, classify the data as being sensitive, based on the definition from the GDPR. For that reason, GDPR obligates to a secure protection of that data and the article says that encryption is needed both for accessing and store that data.
Based on the workflow of the Monitoring System, the authors were able to resume it into four operations, the access, store, profiling and storage. To every operation, were associated the corresponding GDPR rules, that the authors transformed into legal questions. For example, the store operation is reasoned not only by the article 17 of GDPR which is related to the capability of the user to delete their data, but also by the article 5, related to the storage limitation. To the article 17 the authors formulated the question ”Does your device enable users to delete their data in the original used device”. To the article 5 were elaborated the questions, ”How long will the data be stored” and ”How long it is necessary for processing data through your device”.
Moving to the chapter 3, the authors explain how to verify the GDPR rules through the blockchain, based on their abstract model for user privacy.
The first step in the article to verify the compliance, is based on the operations mentioned before. For each one the authors design an algorithm where the input varies accordingly the operation and the output is a boolean value saying if it is in compliance or not and then stores it in the blockchain
In order to give consent to process the data collected, the authors, used smart contracts. This method is done after specifying the compliance of one of the operations and it is retrieved from the blockchain by the data subject, who votes on the operation performed.
Another step in this approach is the contract verification, that aims to analyze if some operation was done without consent or if some of the data was ”processed by the operation are different with those already claimed” by the ”GDPR-operation contract”. In this step the authors introduces another actor, a verifier, which is a ”third party connected to the blockchain”, that votes after any violation was detected, and store this information in the blockchain. The authors in this approach assign the task of report for breaches, that is in the article 33 of GDPR, to the verifier instead of the data controllers.
After the authors demonstrate their results, they conclude that were able, with the introduction of GDPR and blockchain, to demonstrate several design patterns that can be used in IoT since design phase, in order to enforce the processing only under user consent and also to improve privacy in this environment. The use of these patterns not only can be used in IoT but also on cloud computing.
Overall this article introduces a new way of analyzing the GDPR regulation by transforming the articles into a group of questions. The use of blockchain, promotes the transparency required in the GDPR, giving to the data subject the control over their data. Even though, the use of blockchain can increase the resources needed to be performed, the use of this approach without it can also be useful to companies in order to be in compliance.
Personal data related with health is one of the most critical data being collected nowadays, due to the value that it can bring to companies, once for example the pharmaceutical industry is one of the most profitable industries in the world, and also the value that it can bring to the individuals if it is mined using privacy and security methods. So, the article [58] introduces how these systems can be designed and architected in order to be in compliance with GDPR. This article is focused on the topic Healthcare industry 4.0, based on the concept of Industry 4.0 that is justified with ”the increase of digitally networked and data-intensive are pushing forward the smarter production concept and, thus, the industry 4.0 concept”. Based on that, Larrucea et al.

gives as example the National Health Systems (NHS) used across the Europe and being connected through the OpenNCP. Is then identified some problems such as the exchange health records, as well as, vulnerabilities that can be exploited to produce ”unpredicted behaviours”. Besides all those challenges there is also the need to have an evaluation of legal aspects such as the GDPR. So to overcome them Larrucea et al. introduces a healthcare industry 4.0 architectural model, an integration of different tools for assuring security and privacy identifying some of the threats security measures. Then a case study to illustrate these topics.

After the introduction of OpenNCP, GDPR is also introduced, highlighting the consent aspect, which is frequently done in a piece of paper or even worse, there is no record of them. But in healthcare systems is required that consents must be done in an explicit and trusted way, having as base consent management systems.

The architecture proposed for these type systems is based on Reference Architectural Model Industry 4.0 [59], having a structure based on a stack of layers from the physical environment to the digital one. Each one is isolated and responsible for a defined set of tasks, but sharing information between them. For example the Communication layer is responsible for sharing patients' information, which was firstly obtained on the layer Physical Things.

This article then, introduces new tools to be inserted in some of those layers to enhance security and privacy. One of those tools is responsible to hide the data, using a set of requirements as hiding information on the payload or applying anonymization and pseudonimization techniques. This tool is then used on the data layer. Another one is the consent management, used on the business processes layer. This one plays an major role in the structure since the consent in some cases needs to be given remotely. So the need of a trusted framework is needed to provide trust to data subjects. Larrucea et al. approach aims to provide ”integrated set of tools that supports and enables the creation of a formal structure for abstraction, governance and implementation of trust relationships and security policies”.

In order to apply this set of tools, the case study introduced is based on the exchange of information between the United Kingdow NHS and the Spanish one due to some health issue and then there is the need of the Spanish medical staff to access the patient health records for a better treatment. The base here is that the NHS of both countries ”has the same set of tools for managing consent, for hiding sensitive data, and for secure monitoring”. Then, is stated in the article that ”assessment of GDPR within this case study is a complex process”, involving an in-depth analysis of the GDPR regulation, even though, the tools developed responded to the main concerns of each principle. For example the use of OpenNCP to assure the data portability or the tool to hide information answering to the article 5 ”Principles relating to processing of personal data”.

This article approach a very dedicated topic, involving critical data, that must be protected but also available to improve the users life quality. So the authors, introduce a way to achieve that goal in a well structured way, giving specific details about the implementation and in the end complementing them with a real world illustration.

Table 2.3: Comparative between the GDPR articles analyzed

Approach Solution	Introduction GDPR	GDPR Issues	Data Protection Methods	SW Dev. Practices	Solution Introduced
Ayala-Rivera et al. [6]	Addressed	Addressed	Addressed	Addressed	Addressed
Tikkinen-Piri et al. [41]	Addressed	Addressed	N.Addressed	Addressed	N.Addressed
Wachter, Sandra [56]	Addressed	Addressed	Addressed	N.Addressed	N.Addressed
Barati et al. [57]	Addressed	Addressed	Addressed	Addressed	Addressed
Larrucea et al. [58]	Addressed	Addressed	Addressed	Addressed	Addressed

The table 2.3 summarizes the articles analyzed in depth above. The point was to know what are the problems often related with GDPR and by the column GDPR issues is possible to conclude that they exists and are a problem in today's application development. But to overcome them the following columns topics, which were in the majority of the cases addressed, shows that there are solutions and methods to be used and applied.

2.5 GDPR related work

With the need to be in compliance with GDPR, several investigations, work and tools, have been done in order to help and guide companies to achieve it. Besides the ones introduced in the previous section 2.4, it was also gathered specific tools, based on checklists and other methods to demonstrate the GDPR regulation and how to achieve compliance.

Accordingly to this, some solutions based on checklists have been developed, some by public agencies other by privates, such as the [60]. In this tool is made available several grids to help mapping the data in order to identify where remedial actions should be done. The checklist available here [61], splits the regulations into 4 categories, having on each one the articles linked to the specific subcategory being evaluated. The one presented by ICO [62], is more intuitive and clean than the previous one by being truly a checklist. This one is divided again into 4 categories with several questions and also with more information regarding each point, so it is easier to answer. Another GDPR assessment tool is the one introduced here [63]. When comparing with the solutions above, it presents itself, divided into more categories, which improves the user readability. It address topics such as the principles of processing personal data, rights of the data subject or data breaches. Also, inside each topic it clarifies exactly what is the point and its implications, complemented with a link to the GDPR corresponding article. Microsoft also has its own set of tools [64], which covers a lot of aspects by firstly giving a checklist to "simplify GDPR compliance efforts" to a compliance manager, so it is possible to check in real time the risk assessment on Microsoft cloud services and also to give recommendations with step-by-step guidance.

All of these checklist identify the most common regulation obligations, however precise techniques to be applied to each category are not given , in order to be easier for the IT experts to apply them.

On the other hand the solution presented by *Snow Software* [65], which is a paid solution, but does a scan of all the "of all devices, users and applications", highlights what applications can represent a data leak, covering GDPR risk areas by exporting and analyzing the data, among other benefits. However is a paid and closed software as downsides.

In this area of identifying Legal Requirements, Christmann et al. addresses in their article [66], the problem for small companies to access expertise related with these subjects. To solve that, they propose a solution based on a cloud, that is justified due being more flexible and also allowing those companies to save money. The solution consists into identify IT security and legal requirements depending on the functional and non-functional requirements.

Boella et al. do in their article [67] an analyze of the legal requirements in engineering. The goal is to compare existing mechanisms to extract legal requirements and put them in way so that "industry experts" can with more clarity be informed and make judgments on legal requirements. Is then identified a problem among many articles and methods. They often refer the problems, like the ambiguity on regulation, but only a few introduce ways to overcome that issue. So with their solution, this problem would not be the issue that is today.

The solution presented by Gjermundrod et al. [68] gives a concrete solution to address the GDPR data processing requirements. Their solution consists on a PbD framework, grounded by 3 modules, with the possibility to add more modules on top of those. On their solution they give technical information on how to collect the data, how to provide data tractability and also on how to share the data with other entities. All of this being in compliance with GDPR and also providing to the final user the ability to track back their own data.
More recently Tsohou et al. address this problem in their article [69], with the difference of being directly related with GDPR. Firstly is enhanced the advantages that such regulation has in today's world by giving more rights and control to the data subjects on their data. Then, as also mentioned on the above articles, the problem of complexity in order to be GDPR compliant. So, they introduce a new method where is implemented an elicited requirement to help software engineers. However is a work in progress, so it has not reviews.
To complement the articles addressed above Akhigbe et al. [70] questioning the methods being used and the purpose of them to help in the identification of regulatory compliance. So they state that modelling methods mainly focus on "intent of a law", meaning to use it as a guide for be interpreted and applied. But when using "goal-oriented modelling methods", which focus also on how law is structured, it gives in better results. This happens because, they are based on a structure that encompass the requirements in a way that is more likely to fulfil the goals proposed.

Table 2.4: Comparing legal requirements extraction methods

Approach Solution	**Legal Requirements Issues**	**Presented Solution**	**SW engineering methodologies**
Christmann et al. \cite{Christmann}	Addressed	Addressed	Addressed
Boella et al. \cite{boella}	Addressed	Addressed	N. Addressed
Gjermundrod et al. \cite{Gjermundrod}	Addressed	Addressed	Addressed
Tsohou et al. \cite{Tsohou}	Addressed	Addressed	N. Addressed
Akhigbe et al. \cite{akhigbe}	Addressed	N. Addressed	Addressed

As did on other sections, the table 2.4, summarizes the articles and organizes them in a way so is possible so see clearly what was studied. Is possible to conclude that all of them address the problem of issues of requirements, often due to ambiguity or cross reference. The column presented solution then, sees if the authors only presented the issues or also presented a solution to overcome them, which was pointed as one of the problems in some articles. The last column, checks for methodologies suggested to be used. This is an essential point due to some of the problems identified be the difficulty of software engineers to understand the requirements. So if new methodologies are introduced, this problem can be solved and in end have better softwares.

Chapter 3

Privacy and Security : A Theoretical Foundation

3.1 Introduction

In this chapter a review and assessment of privacy and security is done, introducing the basic definitions and terminology and the most common approaches and methodologies used in every application, as well as, the most common vulnerabilities found. Then is then followed by a study of the mechanisms to obtain and increase privacy to the data users and owners. Finally a review of some specific articles which were selected due to their relevance in this subject and because some of them introduce concepts that can be useful to achieve the goals proposed in this dissertation.

3.2 Security of personal data

When it comes to data protection, the goal is to secure the data from unauthorized access and it is sometimes misunderstood or confused with data privacy, that seeks to define who has the access authorization [71]. This can also be seen as, the definition and implementation of technical measures to protect data while privacy is mostly about the law and legal concerns [72], which seeks to define what can be called private.
Privacy is often an ambiguous topic, and as said before, is defined by laws, rather than what individuals consider to be personal [72]. Carrie Gates, introduces this discussion in the following article [72] where data related to data subject such as phone number or addresses are considered ”personally identifiable information”, and so is protected by the companies which collect and store that data. It is also mentioned that financial or medical data is controlled by legislation, which is true. But now, under the GDPR, the personally identifiable information, is also controlled by legislation, by measures such as the need for consent or storage limitation. Also mentioned is the need to have control on the data that is available online, due the fact that people can consider that some type of data is personal to them, and they want to have control to who can see it. This leads to data being available to everyone, which was not intended firstly, due to the lack of control available to the data subject or on a bigger scale, to data leaks. This point is described in the article as, ”fine-grained access control”.
Even though, this type of access mentioned in this article is more related to the possibility of the final user to define who can can see their photos or some particular personal information, when it is analyzed under the security point and also the GDPR, is related to what type of data can be collected and stored without consent and who can have access to that information. So the key to data security and privacy protection issues, is the ability to isolate sensitive data and define the access control [73]. The isolation of sensitive data and privacy data identification, should be seen as primary tasks on a project and considered during the design of applications [73], which is a technical procedure called privacy by design.
The definition of privacy and what it comprises, does not reflect all the problems associated with it. As introduced on section 1.1 ubiquitous computing, as has base the capability of each

devices, due to its computational power, to effectively communicate with another ones. With this is possible to access data and information anytime and anywhere. Ensuring privacy on such devices and communication is a must. Even though there can be several ways to define privacy. Leithardt in [74], describes an approach on how this my be achieved. This one lays on the definition of six groups, each one having its own essential characteristics, which were highlighted by many researchers in this area and mention on the thesis ,to be used when necessary. The author gives as example the User group, saying that it needs ”collaborative so there is interest among others, flexible so that information can be exchanged”. When applying the same method on other device group for example is possible to define privacy over the characteristic of registration or localization. This strategy can be seen as divide and conquer, because if he define the privacy on each group using the characteristics, is possible to build a global definition of what privacy can be.
Also with the introduction of the big-data era, another issues related with privacy are emerging, like profiling [75]. It is identified in this article that some companies believe that after processing data anonymously, the identifiers will be hidden. But what happens in reality is, with that method only, the protection of privacy can not be achieved, because in the end, even carefully handling the data, it is possible to do the re-identifications.
Data protection on the other hand, is a well defined field, it proven approaches and methods, mainly because it is grounded by mathematical proofs, such as the Transport Layer Security (TLS), that implements asymmetric cryptography methods, which has as found the number theory [76]. This protocols and others methods will have an in depth analysis in the sub sections bellow.
Even though data privacy and data protection, are different topics with different approaches, but it shall be used together. This relation can be easily understood because, if data is not well secured and protected it will affect directly the privacy of the data subject.

3.2.1 Encryption

In modern applications, is plausible to say that data follows a stream, starting from its introductions in a system until its output. Along this cycle and depending on how critical and sensitive the data is, the stages may have encryption mechanisms, some of them more adequate to some situations than others. If one of these stages is vulnerable to attacks like MITM or Structued Query Language (SQL) injection, data is no longer protected and consequently GDPR is violated. The first stage where data should be protected, must be since the user is typing in their data, then when it is being communicated over the network to its final destination, for example, if it is done using the HTTP protocol then TLS should be used in order to maintain the data integrity and private or some private communication (VPN/IPSEC etc) must be used. Even when data is kept stored in some server it is important that is kept encrypted, as stated on article 25 of GDPR. This state is called at rest. The other two states are at use and at transit. GDPR requires that when data is at rest or at transit it should be encrypted [77]
With this there this the problem of performance even though applying security mechanisms, like TLS are cheap [77]. Another problem that emerges is if data is kept encrypted, how can data be processed?. The answer to this is to use homomorphic encryption. Craig Gentry [78] define a ”fully homomorphic encryption scheme that keeps data private, but that allows a worker that does not have the secret decryption key to compute any (still encrypted) result of the data”, resulting into enable computation on encrypted data, maintaining the privacy. The problem identified is the excessive computational time it required, which makes this technique

impractical

3.2.2 Privacy Preserving

This subject was introduced in this research by the article, ”The computer for the 21st century: present security & privacy challenges” [14]. The significance of this subject is justified in order to ”ensure democracy and avoid a surveillance society” [14]. To understand this quote, the following example [79], describes the importance of having a balance between the need of privacy, that is often associated with having personal correspondence or to the concept of family, and the need to combat terrorism and fraud. The example is based on the attacks on the 9/11, where after that the government started to collect and store data, such as passenger name records, without be contested. The following image 3.1 complements the example above.

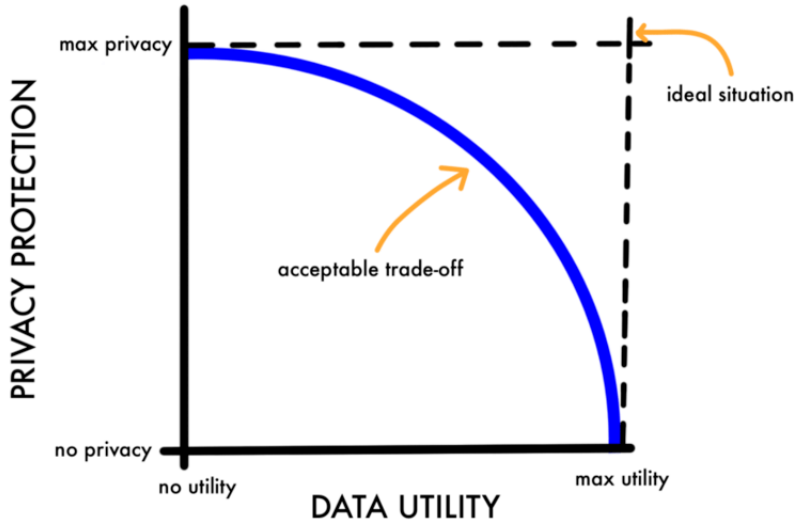


Figure 3.1: Data Privacy Protection versus Data Utility [18]

With the increasing storing and processing of personal data, as well as the data sharing, mentioned on the the previous chapter, from new technologies, the privacy preservation is at stake. Data sharing it is even considered an enemy of privacy preserving [80], which is justified by the wrong use of personal data, that can cause several problems to the data subject, but if it is used for good purposes it can help in the identification of terrorists or help in diagnostic decisions.
So, this subsection is focused on study methods that seeks to hide the information from malicious attackers mostly when it is being processed, mainly because the problem is not when the communication or the authentication is done, but then it is stored [81].
The use of methods such as the Privacy Preserving Data Publishing (PPDP) or the Privacy Preserving Data Mining (PPDM), that respectively, explore methods in order to mask data for publishing and also methods to limit the additional information that can be extracted from published data [80].

3.2.2.1 PPDP

As said before this approach aims to explore methods that seeks to hide or change sensitive information about a specific individual, through methods such as de-identification, that are

present in data sets, in order to provide them to posterior analysis and also to make those data sets less specific, so the data subjects are protected [20]. The importance to study and extract valuable information from these datasets so it can be used for the public level, is at the same level of importance as the need to protect the data subject, because these data sets contain information about healthcare or salaries, so the release of these raw data sets is not correct [82].

Fung et al. [19], divide this method into data collection and publishing, identifying in the collection phase the data publisher, which collects the data from the records owners. On the publishing phase, the data is released to the data recipient, which are the data miners or the public. The image 3.2 shows the mentioned workflow

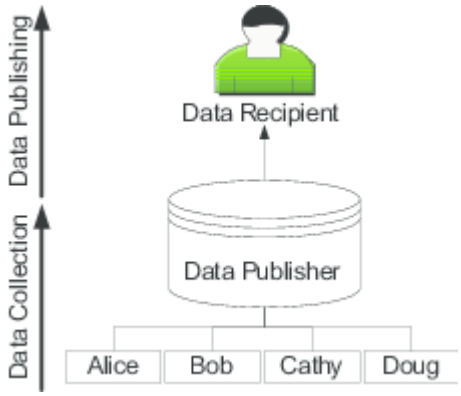


Figure 3.2: Data collection and publishing [19]

Also in the same article,is identified two types of data publishers . The *"untrusted* publisher", which may try to extract sensitive information and the "*trusted* publisher", which the individuals can trust their data, being the untrusted actor the data recipient. Inside the domain of trusted publisher, is was also identified that when the anonymization is not done by experts, doing only the simple tasks of removing the direct identifiers is not enough to protect privacy[20], mainly because when combining these data sets with others, which already identify the individuals, it is possible to re-identify those data subjects[20]. The image 3.3 exemplifies this situation.

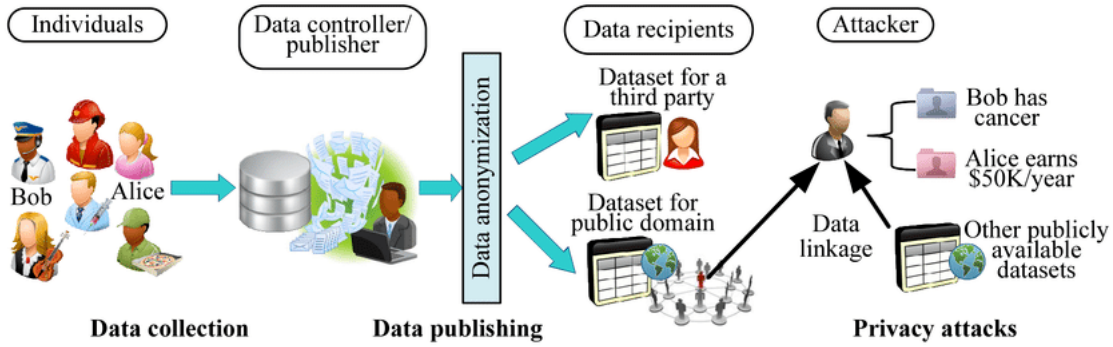


Figure 3.3: Data collection and data publishing [20]

So, the situation that emerges here and based on the figure 3.3 is the need to find a balance between the utility of the data and the these privacy, in order of the data set can be useful but at the same time protect the privacy of the data subjects [82], even though sometimes concerns about privacy can limit the publishing of data [80].

In order to achieve the balance Chen et al. [82], found important to define three components to apply in a data set. These are, the *"Sanitization mechanism"*, that seeks to make data

less precise, using methods such as generalization, that can be for example the grouping of ages under a range or the exclusion of names from a data set. The following component is the *"Privacy criterion"*, defining if a data set is safe or not for publishing using algorithms such as *k-Anonymity*. The last one, is *"Utility metric"*, aiming to quantify the utility of the data set, being most of the time about the information lost on the sanitation process.

Even though this balance is reached and all the security measures were taken to protect the user privacy, a privacy breach can occur. To know when is happens, firstly is important do know what is considered a data breach [80]. In their study Menzies et al. [80], mention that the optimal result from an attack to a data set, happens when the attacker can not learn "learn anything extra about any target victim, compared to no access to the database, even with the presence of any attacker's background knowledge obtained from other sources" [80].

From this definition comes the technique anonymization, which consists into disguise the data so it can be shared for others to extract valuable information without causing breaches in data subjects privacy and will be approached on section 3.2.3.

3.2.2.2 PPDM

After successfully collect or share data, the data mining process happens, allowing to dig out information from where it seemed to not exist anything valuable. Moreover, sharing the results of that process as become a trend among companies, to obtain mutual benefits [83].

The main goal of this process is to modify the original data in order to keep data private and also the knowledge private after the data mining process [84]. To ensure that privacy is not threaten Verykios et al. [84] point out 3 techniques. The first one is based on heuristic methods which seeks to modify only selected values in order to "minimize the utility loss rather than all available values". The second one relies on cryptography and ensures that in the "end of the computation, no party knows anything except its own input and the results". Finally the third one based on reconstruction, aiming to rebuild the "original distribution of the data from the randomized data".

Even though it is a similar field related to PPDP, is has some limitations [85] and in the same article is pointed those limitations, such as it is more focused techniques of publishing data and not on techniques for data mining and also it does not prevent the truthfulness of data. This last limitation is justified with the need to connect the data with a data subject. Fung et al. [19] states the following example "The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug". If the result of this mining does not relate to an individual is may be difficult to deploy the result to the real world. As said before the cryptography is one of the methods inside this area, but also is not useful since it "hides the semantics required for acting on the represented patient" [19], although it connects to an data subject.

3.2.3 Anonymization and Pseudonymization

This process, as studied before, encompass a various number of techniques in order to make data anonymous and more important, to protect the user's privacy. Also, it is a process used inside the PPDP in order to share data for posterior analysis. By saying that data is anonymous, is meant that the data subject is no longer identifiable, making this an "irreversible" process [86]. Inside this scope the protection laws are not applied [87].

Under the GDPR law, more precisely on article 6, ”Lawfulness of processing” [88], it is mention the use of pseudonymization as one of the ”appropriate safeguards” as well as encryption. Pseudonymization is then defined as ”means 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 organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person” [89]. From this definition is possible to state that personal data in which was performed some pseudonymization techniques, are still under the GDPR regulation, because it is personal data, even though it is called pseudonymised data. Those techniques are based on replace the direct identifiers with a pseudonym, for example a random generated number and then on a new table stored in a different place, store the linking between the pseudonym and the identifiers. The image 3.4 describes this process. More precisely, these techniques can be divided into two categories depending on how the

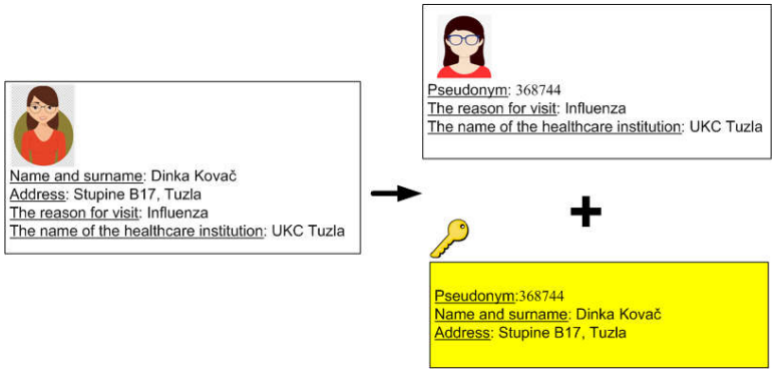


Figure 3.4: Data Pseudonymization [21]

pseudonym is generated [21]. The first one encompass ”techniques in which the pseudonym is independent of the original data – e.g. tokenization” and the second based on ”techniques in which a pseudonym is generated from source data e.g. pseudonymization by encryption or pseudonymization by a hash function”.

The question now, is how to choose the best technique. In order to answer that, the Guideline ”PSEUDONYMISATION TECHNIQUES AND BEST PRACTICES” from ENISA [90], says that the choose of a Pseudonymization technique should be based on the appropriate protection level and in the utility of the pseudonymised dataset. Regarding the protection level, that guideline refers the use of Random Number Generator (RNG), which is an example of tokenization, the use of Media Access Control (MAC) as long as the key can not be compromised and also the use of encryption algorithms such as Advanced Encryption Standard (AES). It is also stated on that guideline that for data to be useful, the entities responsible for performing for applying Pseudonymization, might use a combination of the techniques mentioned or a varition of a selected technique.

Taking into account that Pseudonymization is a reversible technique, anonymization on the other hand, should be an irreversible technique. Sophie et al. [87] conclude that a zero risk is not possible to achieve on anonymized dataset and also based on the following study [91] is was possible to re-identity 99.98% of the data subjects. With that being said, it is concluded by the Opinion on Anonymisation Techniques [92] ”that anonymisation techniques can provide privacy guarantees and may be used to generate efficient anonymisation processes, but only if their application is engineered appropriately”, meaning that the optimal solution should be build on ”case-by-case basis” and combining a set of different techniques. Those techniques are also mentioned on that Opinion as being the following, noise addition, permutation, differential

privacy, aggregation, k-anonymity, l-diversity and t-closeness.
Before analysing those techniques is important to define the attributes that are present on a data set [17]. Those attributes are, usually, the Explicit identifier, the Quasi-identifiers (QID), the sensitive information and the non-sensitive information. Also Fung et al. [19] define an anonymous table with the following expression ”T(QID’, Sensitive Attributes, Non-Sensitive Attributes)” where the QID’ is the anonymous version of the original QID, resulting from applying anonymization techniques.
Based on that attributes is now possible to understand and study some of the techniques mention above. To approach the topic of anonymization, it is possible to split it into two different categories. The first one is randomization and the second is the on generalization [92]. Randomization is defined as a set of of techniques to remove the strong link between the data and the individual and can be combined with generalization to ensure the individual’s privacy. Inside this category is defined the use of noise addition, which consists into change the value of an attribute so the value are no longer precise. It is also defined the use of permutation, and as the name suggests, it consists in the shuffling of some attributes. Moving to the generalization, consists in”replacing the values of an attribute with less specific but consistent values” [20], for example instead of writing a birth date as *dd.mm.yyyy*, only write the year or instead of writing a specif value for weight or height, write it as a range. Inside this category is possible to find algorithms such as k-anonymity, which assures that each record in a dataset is indistinguishable from at least k - 1 [15] other records and to achieve this values are generalized or aggregated.
These to terms, Anonymization and Pseudonymization, can be confused by being the same. In fact, Pseudonymization is not even a sub category of Anonymization, but instead a security measure [93]. The point is that Pseudonymization can help in the anonymization process when used with other techniques such as generalization and deletion of data [94], but when used alone can serve to the lawfulness of processing [94]. Also, as Bolognini et al. [93] mention, it may not be possible to make data anonymous and at the same time keep it useful, so when Pseudonymization is used, it not only can serve to protect the user’s identity, by removing it but also allows to do the re-identifying when needed. Furthermore, is helps to ”fulfill the obligations of data protection, especially in terms of accountability”.

3.2.4 Privacy by design & by default

The importance to study this topic comes from the fact that, under the GDPR regulation, the data controller must implement this topic, as mentioned on the Article 25, in order to ensure that ”only personal data which are necessary for each specific purpose of the processing are processed”, as well as one of the techniques to implement the appropriate technical and organisational measures. Also it can serve as proof to demonstrate accountability.
Even though is mentioned on the regulation, the PbD should already be a standard for companies that deal ”with strong privacy policy and take data breaches into account when building new systems” [95], however for start-ups it can be a problem [96], but on the long rung it will bring to the company a good reputation on how they handle the data. With that being said, a good implementation of this topic depends on the knowledge that the ones who are going to develop a system, have on privacy [97], so the work related to the data privacy and protection should be supported by a privacy specialist. Besides that, Yod-Samuel Martín et al. [55] mention that PbD ”has not yet gained widespread, active adoption in the engineering practice, due to a mismatch between the legal and the technological mindsets”, which can result from, view privacy in a perspective of data security and also from not take into account data privacy and protection

since the design phase.
To better understand the approach to take on PbD, Cavoukian [98], defined a set of seven principles ”to serve as a reference framework and may be used for developing more detailed criteria for application and audit/verification purposes”. Even though all of them are very important and useful for a deeper analysis of this topic, it is important to mention the second principle, ”Privacy as the Default”, consisting in some of the GDPR principles, such as Purpose Specification or the Collection Limitation which as part of the consent principle. Also the third principle, ”Privacy Embedded into Design”, meaning that privacy should not be considered as an add-on of systems, but as a concept to have in consideration since the system’s design and also as a essential component of the core functionality. The last principle being mentioned is the fifth, ”End-to-End Security”, which can be translated into having security mechanisms to protect the data life cycle, highlighting the point that ”without strong security, there can be no privacy”.
Has mention before, is form the design phase that privacy and data protection should be take into consideration. To help with that, the ”Privacy and Data Protection by Design – from policy to engineering” [99] published by European Union Agency for Network and Information Security (ENISA), mention the use of design patterns, which are useful for making decisions about the organisation of a software system. Design patterns is then defined as ”scheme for refining the subsystems or components of a software system, or the relationships between them. It describes a commonly recurring structure of communicating components that solves a general design problem within a particular context”. With the use of this concept it is possible to break down a bigger problem into several pieces, making the system more manageable, by describes the reasons of the break down. One of those design patterns is the Model-View-Controller as it divides the definition of the data from the view and also from the interaction with the user.
As pointed, PbD can bring to software some advantages and also help to have the necessary accountability required by the GDPR. However, there are some limits to it. As mentioned here [99], the problems are based on the fragility of privacy properties, mainly because systems nowadays are often connected to each other, making hard to assess if that privacy property is preserved in all of them. The limitations are also grounded by the increased complexity of the system even when adding a naive functionality, if on top of that there is the need to implement privacy properties, it will bring to the system a new level of complexity.

3.2.5 OWASP

Since the EPOS GNSS project uses in its architecture APIs and web applications, as mention on section 1.2, to allow users to access data and do operations on it, and since one of the goals is to perform and give in the final report, a security analyzes, OWASP comes in highlighting the most common vulnerabilities, submitted by a several number of companies, specialized in application security[27].
Even though inside OWASP exist different projects, this study will be focusing the Top 10 project. This one, selects the main vulnerabilities found, accordingly to their level off exploitability, detectability, and impact.
Besides those ten vulnerabilities, it is important to keep track on all the others that exists, for example using the OWASP Cheat Sheet Series. The level of effort put into this musk come from the risk measurement, associated with the application. The OWASP authors, mention that the risk depend on the probability associated with each ”threat agent, attack vector, and security weakness and combine it with an estimate of the technical and business impact to your

organization", summarizing it on a table and giving it a score depending of how easy it is exploit the vulnerability in cause, of how prevalent they exist, of how it is easy to identify or not and also of the impacts that such vulnerability can do. It is also based on this table, that the 10 vulnerabilities are classified and also it is given a guide on how to prevent those vulnerabilities from happen.

Going through the list, that will not be covered in its totality, firstly comes the Injection, often through SQL, where data is sent to a code interpreter, deceiving it, in order to obtain data without authorization. To prevent it it can be used parameterized queries or prepared statements, even though, is still susceptible to attacks, so it is also recommend to use of validation mechanism to sanitize the data being sent to de server.

Within Injection vulnerability, is important to highlight the attack Blind SQL injection, because is a parameter on the tools used for the security assessment. This one aims to exploit applications that show generic messages, where the attacker asks the database True of False questions, to achieve his goal. This attack is similar to the traditional SQL injection, but no data is retrieved.

The second on the list, refers to Broken Authentication, happening when authentication mechanisms are not correctly implemented, which allows the attackers to access other user's identity. The use of multi-factor authentication, use of weak password mechanisms or an increasing delay between failed login attempts, are measures used to overcome this issue.

Going to the third vulnerability, the Sensitive Data Exposure, which as the name suggests, refers to the lack of protection over sensitive data, such as, healthcare data or financial data, which can lead to attackers to commit crimes in name of others. To avoid this exposure, it is needed to classify data depending on its sensibility and accordingly to the privacy laws. Also the encryption of all sensitive data, as well as, the encryption of all data in transit using TLS.

The next vulnerability, is the Broken Access Control, which allows attackers to bypass the authentication mechanisms in order to have more privileges and access to unauthorized functionalities or data. The use of authorization tokens, so every request require that the authorization token be present [10].

The last but one addressed, refers to Cross-site scripting (XSS) attacks, that the goal is to inject malicious scripts into trusted websites. This is often done by a web application to send malicious code, in form of JavaScript, to the end user, which will execute the script unaware of the danger, because it is embeded on the source code.

Inside the XSS there are three types of attacks used. The first is *Reflective XSS*, where the victim clicks on a malicious link, and the result of the script injected is reflected on the victims browser, for example redirecting him to the attacker site stealing the victim's cookies. The second attack *Persistent XSS*, is similar to the first, but now, the attack is affects directly everyone who uses the site, because the malicious script is stored on the site or even in a database and then is loaded into the victim page, because again there is no control to what is being rendered. This happens due to lack of input sanitizing. The third type is *DOM Based XSS* and the goal is to attack the data being written in the Document Object Model (DOM), that does not have proper sanitation There are many ways in which an attacker can entice a victim into initiating a reflective XSS request. For example, the attacker could send the victim a misleading email with a link containing malicious JavaScript. If the victim clicks on the link, the HTTP request is initiated from the victim's browser and sent to the vulnerable web application. The malicious JavaScript is then reflected back to the victim's browser, where it is executed in the context of the victim user's session.

The last vulnerability mentioned here, refers to the insufficient logging and monitoring, which allows attackers to further attack systems and tamper, extract, or destroy data. To prevent

this is necessary the implementation of measures, such as, the generation of logs that can be read by a log centralized solutions. Also use of integrity controls, allowing only the appending of data and not the deletion or modification of the records.
The use of this guide in combination with other tools, are a good starting point in security assessments, even though, it was not covered in its totality, is highly important to have present those 10 vulnerabilities as well as others that exists or are emerging. So the need to have an up to date and automatic system that analyzes this issues are extremely important for companies to be more secure and at the same time being in compliance with GDPR, because there is no data privacy without security.

3.3 Article analysis

Data privacy and security has become the object of a large number of studies, not only because of its impact in business in terms of data as an asset, but also because of the many challenges that it introduces. This is so, because with the emergence of new terms such as big data and IoT, companies now have as an asset the data that users trust them with, so their expectations are that companies will not jeopardize this data.

- In the article ”The computer for the 21st century: present security & privacy challenges” [14], Oliveira et al. review the security and privacy subject in *UbiComp* systems. Since this paper is focused into those systems, which comprise Big Data and IoT, is extremely important to identify and study the issues associated with it. The authors identify seven areas, each one, covering different challenges of privacy and security. They start with the use of weakly typed languages, used by some resource-constrained devices, going then to the second area identified as Long-term security, where they analyze some cryptosystems based on Rivest-Shamir-Adleman (RSA) or Elliptic Curve Cryptography (ECC).

  Studying this area in more detail, it is stated from the beginning the importance of having systems that can be reliable for years because of the investment that is done, so if it stays usable during a longer period of time, it will give a better income. When this situation arises from *UbiComp* systems, the solution is to assure that, from the design phase to the latter phases of development, security needs to be taken into consideration, in order to reduce the need of future updates, which increases the costs of development, not only because it can open new windows to security breaches, but also because of the incessant emerging of new devices.

  It is also mentioned that in can be very demanding achieving long-term security, mainly because it requires that each device must be future-proof. So, as the authors refer, most of the cryptographic techniques, depend on the cost of computational problems, which based on that, they are considered secured. The problem is, to achieve the future-proof condition, these techniques can be useless. The first threat to it, are the advances that are being done in order to solve the problem in a more efficient way, which will consume less processing time. Another threat is the development of quantum computers, that, as the authors show on a resume table, can be used to in a efficient way, crack symmetric cryptography, being the solution the increase of the keys size, but keeping in consideration the performance needed to do the calculations.

  The next relevant field, is the cryptographic engineering. This area was introduced on the previous chapter since this one, is identified in this paper as the ”main ingredient of most

security mechanisms". The authors begin this chapter by mentioning the various parts, being it hardware sensors, communication protocols or human factors, that constitute the core of a *UbiComp* system. The result of such a complex system, is that the attack area to be explored is larger than the ones in "traditional computing", so the solution is to have a combination of techniques to protect those systems.

Since cryptography is such a fundamental component in any system, the authors identify it, as not only the weakest component to be compromised. Once it is compromised, it can have severe consequences, such as data breaches or affect the availability of the system.

The remaining of this chapter, approaches distinct areas inside the main one, the cryptography. For example, it explains what is lightweight cryptography, which is the design of "application-tailored cryptography" mechanisms, in order to reduce the consumption of energy or the number of processor cycles, so with can be used in "resource constraint devices". Also related with the resource constraint problem Saraiva et al. [100] does an benchmark of some encryption algorithms such as the AES with different keys lengths or SPECK, which is considered a lightweight block ciphers , in order to observe how much time and how much energy it was needed to apply the algorithms in files with different sizes. This benchmark was done on two smartphones and the results were that it mainly depends on the CPU capacity and that AES is the most time and battery consuming.

The next important chapter of this article is named resilience and it is defined by the authors as on of the foundations of security. The justification to that assertion, is due to the fact of having the capability to recover or mitigate damages and also financial loss that can result from the some service being unavailable. So the goal is to "identify, preventing, detecting and responding to technical failures". Also identified was the growth of intentional failures, resulting from exploiting services, mostly caused by Distributed Denial-of-Service (DDoS). DDoS attacks are caused by using a huge number of infected internet connected devices, such as printers or IP cameras. This is where resilience comes in, giving the ability to mitigate those attacks and swiftly restore the availability of those services.

Moving onto the the seventh chapter, the topic of privacy implications. Here the authors start by highlighting the huge quantity of data generated by **UbiComp** systems and the importance for companies and also for society to be able to extract benefits from it. With these advantages also comes the other side, which is data being used for illice purposes. In order to have privacy, is mandatory to establish security, but the opposite is not true. The example that the authors transcribes exactly what happens. If there is a communication service between a user and service provider, if that communication is not secure, it is not possible to ensure privacy. But if the communication is secure, it does not mean that is private, because the data transmitted can be used in ways that is should not be. So, the writers state that the first step is to find the extent of the data and the impact of data leakage.

As sub chapters, the authors enumerate several problems related with privacy, being the first one the need to classify what type of data is sensitive or not, mainly because the definition of what is private or not may differ from country to country. The other issue, is the large amount of laws, which leads to need to go through a lot of bureaucracy. But also the lack of laws, can be a problem, causing some companies, which are not the most ethical or are headquartered is countries that do not have a good regulation, to have advantages over other companies, on how they use the private data. The solution given is

to use the privacy-preserving techniques to ensure democracy.

The following sub chapters of the seventh chapter, introduce data anonymization and how to measure the degree of anonymity, using several methods. Since it is an important topic but also very technical, it will be approach on section 3.2.3. It also introduces the use of homomorphic encryption in order to ensure privacy-preserving systems, but also the problems regarding the performance of using such type of technique.

This article ends with the conclusion, where is important to enhance the challenges that these types of systems are bringing to investigators and what are the main topics inside privacy and security, which can serve as a guide in the opinion of the authors, to future investigation.

- Since most of the data collected nowadays, is stored in cloud architectures and is used to the processing, is important ensure that them, are aware of the security needed, in order to protect the user's privacy. With this, the paper "Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures" [101], presents a set of tamper-proof security protocols based of cryptography and also with this type of service, gives to the users, more control over their own data,

  The authors of this paper start by enumerating some of the advantages of use cloud computing. It brings to companies a layer of abstraction, because the storage and also the processing, is done in remote computers instead of using their local ones, saving them the need to worry about physical resources. It also brings to companies an another level of commodity by enabling "elastically", which can give a promptly response to the users needs, due to virtualization techniques.

  With all the migration that is happening, due to all the advantages, there are several present and emerging challenges. The authors mention that the most important of all, are the privacy and the security topics. Following that line, the authors also state that, privacy should be provided to customers at a minimal cost and that it should be made available in order to be more configurable and also user-friendly.

  As solution to this issue, the authors suggest *PasS*, which will be reviewed, in order to extract information, that can be helpful in the remaining of this master thesis.

  In section three the authors define what is a cryptographic coprocessors, which in this solution provides a secure and trusted environment in cloud computing. In our study, the subsection of this chapter is not completely relevant, but is important, in order to have an overview of the solution.

  So the description of cryptographic coprocessors is stated by being an "small hardware card that interfaces with a main computer or server, mainly through a PCI-based interface". The advantage to use this card, by the optics of the authors, results from is it tamper-proof casing, that is resistant to physical attacks and when it detects an suspicious physical activity it can reset the RAM, persistent storage, processor registers.

  Since it is such an critical component in the cloud based system, it should by provided by a trustful third party entity and it should be installed in every physical server running a virtual machine. Also to make it more affordable, this service can be shared across all the users of the cloud service.

  The authors highlight this trusted third party, as the main factor of having this shared service, because it is responsible to load a set of public and private keys associated to

each customer, to a persistent storage present on the cryptographic coprocessors, being again important when registering new customers. This trusted third party also has its own set of keys in order to authenticate itself on the crypto coprocessor to securely execute commands against it.

On the software layer, the solution presented gives to the cloud customer the ability to configure its own applications to support the security required by the *PasS*, adopting the concept of software division. From the point of view of the authors, this concept allows the customers to classify which are the components that are protected or unprotected, being the protected ones, able to run on the address space of the crypto coprocessor and the others as a common process on the main server.

Also in the chapter three, the authors dedicated a sub section to the topic, data privacy specification. Here is proposed a data classification based on the significance and sensitivity, resulting in three categories, that is done before sending the data to be stored on the cloud. This classification is done by the cloud customer and not by some automatic classifier. The first category has the name *No Privacy* and as the name suggest, the data is not sensitive, so it can be stored without any encryption but if the customer needs, it can use SSL session to send the data. The second category named, *Privacy with Trusted Provider*, data will be stored encrypted, but using an specific provider key. So the customer trusts the provider its own data, encrypted with the provider key. Also the communication is done using a SSL connection to provide data confidentiality. Finally in the third category, *Privacy with Non-Trusted Provider*, where the data is encrypted on the customer side, using a customer specific key, that is shared with the crypto coprocessor, being then stored encrypted, not allowing the cloud provider the access or visualization. This type of data can only be processed in the address space of crypto coprocessor.

The fourth chapter specify the all steps that the customers need to perform, to add privacy measures to their data and software and also how the *PasS* should behave depending on the inputs given by the user. Even though it is divided into three sub sections, which are important, but not highly relevant to our study, it is important to highlight the privacy feedback. This protocol should be considered when designing cloud services, as it shows to the users, what are the privacy methods which are being used on their data and also inform of the risks that a possible leak can cause. The use of this protocol is then defined on the remaining of this sub-section.

This paper shows that the privacy of the data should be fully configurable by their owner, who is responsible to decide what types of data should be considerate private or not. Then, with help of such service, the user can have an overview of how their data is being protected, which transmit security to the final user, for being fully transparent.

- With the increasing number of data being stored in cloud architectures, due to the higher number of data sources available nowadays and also due to the importance of processing data as an valuable asset to companies, the term big data, emerges bringing with it, privacy issues and also the need to protect data. So the following paper entitled ”Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR” [17], brings to the discussion, the impacts that the GDPR will bring to this industry that keeps on rising.

  Right in the introduction chapter, Gruschka et al. state the example, that a supermarket doing targeted advertisement, resulting from profiling techniques and shopping patterns, were able to conclude that a girl was pregnant, which is a privacy issue. These risks does

not stop where, and keep becoming more common in critical areas such as the healthcare or financial areas.

Another problem identified in this chapter, is the effort that companies need to do in the design phase, in order to apply technical measures for privacy-preserving data processing, which some of them are not willing to do, because as the authors state, it can affect the performance of the system.

Moving to the first subsection of the second chapter, where the authors do a review over the GDPR regulation, pointing out some of the most important and relevant articles, they also have a subsection dedicated to technical aspects, which are the implementations of the requirements of the previous sub section, presenting some methods for privacy-preserving data mining. Those methods are the data anonymization and the use of mining algorithms over the data anonymized. In order to support the anonymization the authors classify the data set into four categories. The explicit identifiers are the ones where data is directly linked to an data subject, the quasi identifiers are the ones that can be used to re-identify a individual when used with other data set but can not identify one directly. The third category are the attributes that the data subject does not want to be revealed and lastly the attributes which are not sensitive. In order to measure the level of anonymity the authors point out some algorithms that can be used. Also on this subsection the authors go through a few common anonymization methods like the suppression or the generalization. Since these are very technical and major techniques, a deeper approach is done on the subsection 3.2.3 of this dissertation.

In the next section the authors analyze the use of GDPR with real life research projects, which can be useful to study, in order to see the trade-offs and the implications of such regulation and also to analyze techniques used in a real life context.

The first case study mentioned, develops methods to do analysis into big data, using machine learning and subjective logic, and at the same time it needs to be in compliance with GDPR, so when dealing with fully anonymized data sets, there can not exist any linkage to a specific data subject. One of the datasets that they deal it is the *Sysmon*, which is a a Windows service, that monitors and logs the activity of Windows workstations. Such dataset contains multiple sensitive identifiers, like the accounts username, the IPs addresses, the running process or the internet activity. The problem here is that, one user can be directly or indirectly identified. If they apply anonymization techniques over this data, it would deny the re-searching, leading to use of more complex approaches.

The authors also highlight, how they managed the data storage and the accessibility. So the data stored can only be accessed by authorized researchers combined with the principle of least privilege, which means that the user can only have the essential privileges to execute his function. Also is pointed that the access to the server can only be done from the inside network, which is restricted to a list of MAC addresses, being also protected by a firewall, only allowing connections to the Secure Shell (SSH) port. Also the access to the server is restricted to only the duration of the project, which means that after, the access will be drooped.

Finally the authors analyze the ”Trade-off between Security, Reproducibility and Dataset availability”. The use of Reproducibility, refers to the transfer of knowledge, so others can use the datasets and also the software code. After in stated the use of the Sysmon dataset, even if it is anonymized, to identify possible attacks or vulnerable applications.

The next case study approached in this article is the *SWAN*, that is a project responsible to develop authentication technologies, using biometric identifiers, which are a type of personal data, inserted into a special categories of personal data of GDPR, for which it is required a explicit consent.

In order to be in compliance, the *SWAN* team created a privacy policy, that includes the sections, ”Defintion of biometric identifier and biometric information”, the section of Consent, the Disclosure, the Storage and the Retention Schedule. With this policy the data subjects are informed about the purposes of the data collection.

Also the creation of the data storage was done following the GDPR recommendations, like the use of pseudonymization, allowing the re-identifition of data subjects when required. Following to this, the authors state a few advantages of using such technique. The use of pseudo IDs can faciliate the destruction of the data, when the data subject withdraws is consent, which can be done at anytime. Also it allows to reduce ability to extract valuable information if a data leaks occurs. It is also mentioned the analysis of such datasets, without the need to access the raw data.

Moving to the fourth chapter, the authors have identified the privacy methods required to be in compliance with the GDPR on the *SWAN* project, since it uses critical type of data. The first one was explicit consent, since they were using biometric data. The following one is the security of the processing system by only allowing authorized access and also encryption techniques. The third method is the use of pseudonymization disabling the re-identification. The fourth is the processing using biometric templates, like the cancellation template, which ”allow the revocation of a compromised biometric template”, making re-identification much harder. The last one is the limitation of data storage by defining a maximum period for storing the data and when it is over, the data is deleted.

It is important to highlight the authors observation that, even with the use of anonymization the re-identification is possible, even with all the databases fulfill with the GDPR, that does not have a strong formal definition of this technique.

This article does a good overview of all the relevant articles from the GDPR and with the study cases mentioned, it is possible to extract what are some of the changes that companies need to do to be in compliance, specifically when using such type of data, this article can be a good starting point for some of them.

Table 3.1: Compartive of security and privacy related articles

Approach Solution	**Privacy Isssues**	**Security Issues**	**Cryptography Techniques**	**Privacy Techniques**	**SW Development Practices**
Oliveira et al. [14],	Addressed	Addressed	Addressed	Addressed	Addressed
Itani et al. [101]	N. Addressed	N. Addressed	Addressed	Addressed	Addressed
Gruschka et al. [17]	Addressed	Addressed	N. Addressed	Addressed	Addressed
Saraiva et al. [100]	Addressed	Addressed	Addressed	Addressed	N. Addressed
Gates, Carrie [72]	Addressed	N. Addressed	N. Addressed	Addressed	N. Addressed

The table 3.1 summarizes the articles analyzed, in order to have an overview of what was approached in some articles and missing on others. It is based on the words *Addressed* or *N. addressed*, which stands for not addressed, varying if an article works with the topics used on the table header.

Is possible to conclude that privacy issues in most of the case is a topic addressed, which reflects the concerns about it and why it is a trending topic. To complement this issues the column privacy techniques was used to check if the articles that addressed the privacy issues also introduce the techniques to overcome them.
The same approach was used on the columns security issues and cryptography techniques, even though cryptography are not he only method to surpass security. The column Software Development Practices, as inserted in the table since having them, helps in the design phase, which can increase the application complexity but can save time in the future.

Chapter 4

Application Development

4.1 Introduction

In this chapter a detailed guide on how PADRES was built will be given, together with the justifications of each technology used. By the end of this chapter, the reader should have a good overview of how everything works and also how to use the application.

4.2 Requirements

As identified in the previous chapters, some of the problems regarding the GDPR regulation comes from the difficulty in extracting and applying the law. In order to aid this process, the regulation was split into seven principles, as mentioned in [6], here [102] and discussed in chapter 2. Inside of each one of those principles it was summarized all the points linked to them, in order to show only the essential information, as well as the main points of each one.
By providing the regulation organized in this format, it is easier for companies to know and study the regulation. Besides those principles a new one was added - a country specific requirements. This is necessary to deal with country specific regulations that fall outside the scope of the EU's GDPR. Some countries have their own laws, regarding data protection and privacy which must also be considered in order to have a complete overview of everything that needs to be analysed.
Also accordingly with the previous research, it was possible to state that data privacy without security can not be ensured. So, in order to meet that goal, it was provided a security analysis based on open source tools, such as the OWASP ZAP. The point here is not to replace the security analysis done , but to provide an overview, based on those tools, on how the system is or can be vulnerable.
Once the previous steps are concluded, it will be given a classification. This, is only based on the eight principles, having each one a classification, contributing to a final classification, that will be provided on the report, including several suggestions, to increase the classifications.
With this approach it is expected that companies using this, can overcome the difficulties observed, which were also studied before.

4.3 PADRES Introduction and Architecture

PADRES was designed to be available only locally. Therefore the user of the program only has to install the application on their computer and interact with it through the browser. Other types of deployment are possible, such as being made available online although due to the nature of the application this would require adding more complexity to the application, such as adding logins and user profiles.

PADRES is based on a client-server model, where the client is the front-end, the browser, and the server the backend. The backend was developed based on the REST architecture, using an API to communicate with services without having to know how they're implemented, simplifying the application development, making it an REST API.
REST was introduced by Roy Thomas Fielding in his PhD dissertation [103] to answer problems related with the increasing use of the Web to do business, which existing architectures ould not answer to to their limitations in scalability and extensibility. The definition of REST as given by the creator is "a set of architectural constraints that, when applied as a whole, emphasizes component scalability interactions, generality of interfaces, independent deployment of components, and intermediary components to reduce interaction latency, enforce security, and encapsulate legacy systems.".
The REST is used over the HTTP or Hyper Text Transfer Protocol Secure (HTTPS) protocol, using the HTTP verbs [104] to indicate the action to be used on the server, making this the uniform interface needed to interact between the client and the server, as defined above. Some of the most verbs used are the *GET*, *POST*, *PUT*, *DELETE*. Linked to this, are the HTTP response status codes [105], to help the client understand the response.
The front end was developed using the Angular framework. Angular was developed by google, to help building Single-page application (SPA) and uses TypeScript as its main programming language. As browsers can not execute TypeScript code it is first transpiled by the framework into JavaScript before deployment.
To have a basic understand of what was done in the front end section 4.6, firstly an overview on the basic concepts, in which Angular is grounded, is need. So the the diagram 4.1 put that aspect into perspective, by showing how these concepts are related.

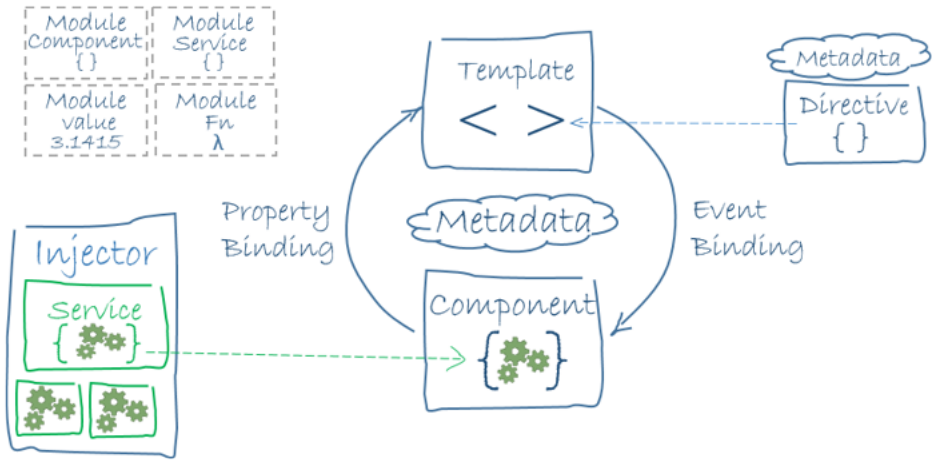


Figure 4.1: Angular architecture [22]

One of the advantages of using Angular comes from its modularity, its own system of modularity is called*NgModules*. These, can be seen as containers, associating components with related code, such as services or libraries. This means that, every Angular applications must have at least one *NgModule*, responsible for launching the application. However the use of multiple *NgModule* is recommended in order to improve a projects organization and modularity, taking advantage of the import and export functionalities.
Inside the *NgModule* there is a parameter called declaration, where the components and directives are declared. Like the *NgModule*, there must be at least one component, the root component. Inside each component the logic is declared and also the data to be used. Also, the

corresponding template is declared, combining the HyperText Markup Language (HTML) with the directives, responsible to define the logic inside the template, allowing to change the DOM. The connection between template and component, is achieved through two-way data binding, meaning that changes in the component are reflected in the template and changes in the template are reflected in the component.
Also inside the *NgModule* the services are declared. These can be injected inside the component, allowing it to access the services, that can be used to declare data structures, to do HTTP requests or just to send data between components.
This diagram 4.1 is based on a Model-view-viewmodel (MVVM) pattern, that has similarities with the Model-View-Controller (MVC) pattern, in which AngularJS was based on. Both share the Model and View aspects, being the difference on the controller, that in MVVM is the ViewModel. This change, allows this aspect to not only interact with the DOM, but also listen to interactions in the view and change it. This difference is achieved in Angular through the two-way data binding.
Since the solution also encompass the use of a database to provide data and also store it, SQLite was used to provide those functionalities, because it is a small and fast SQL database engine, that fits exactly in the proposal solution.
The image 4.2 describes the architecture of our approach.

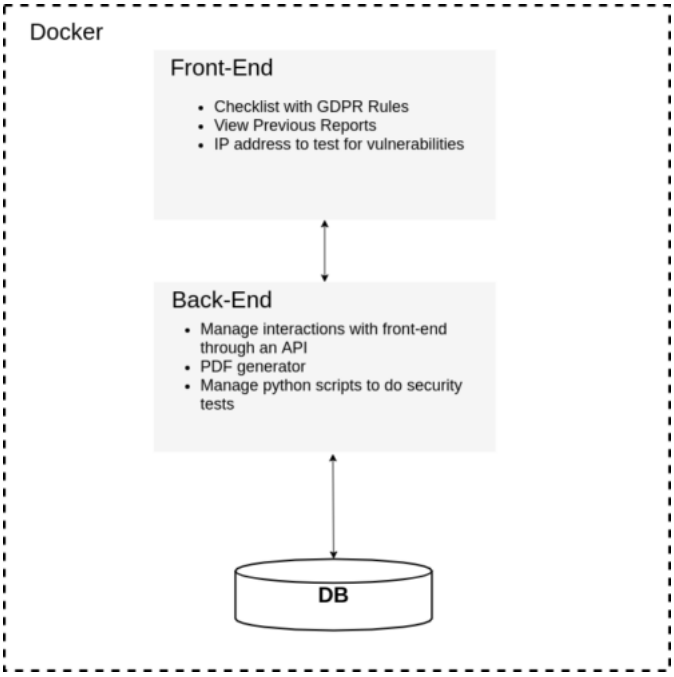


Figure 4.2: Application architecture

With this solution is possible to have different clients interacting with the application, accessing the data provided by the database through the REST API on the browser, giving a easier way to analyse the GDPR compliance. Also, it is possible to add more rules to each principle, which is not provided since it is not needed for this solution and also allows PADRES to be more extensible and scalable and eventually making it available online.
The image 4.3 shows the database schema that is used to support both the frontend and backend. From this schema it is possible to observe that the tables *principleHeader*, *principleID* and *suggestion*, do not have any relation with the other tables. This is because those tables only store information about the principles and do not interfere with the remaining tables. The

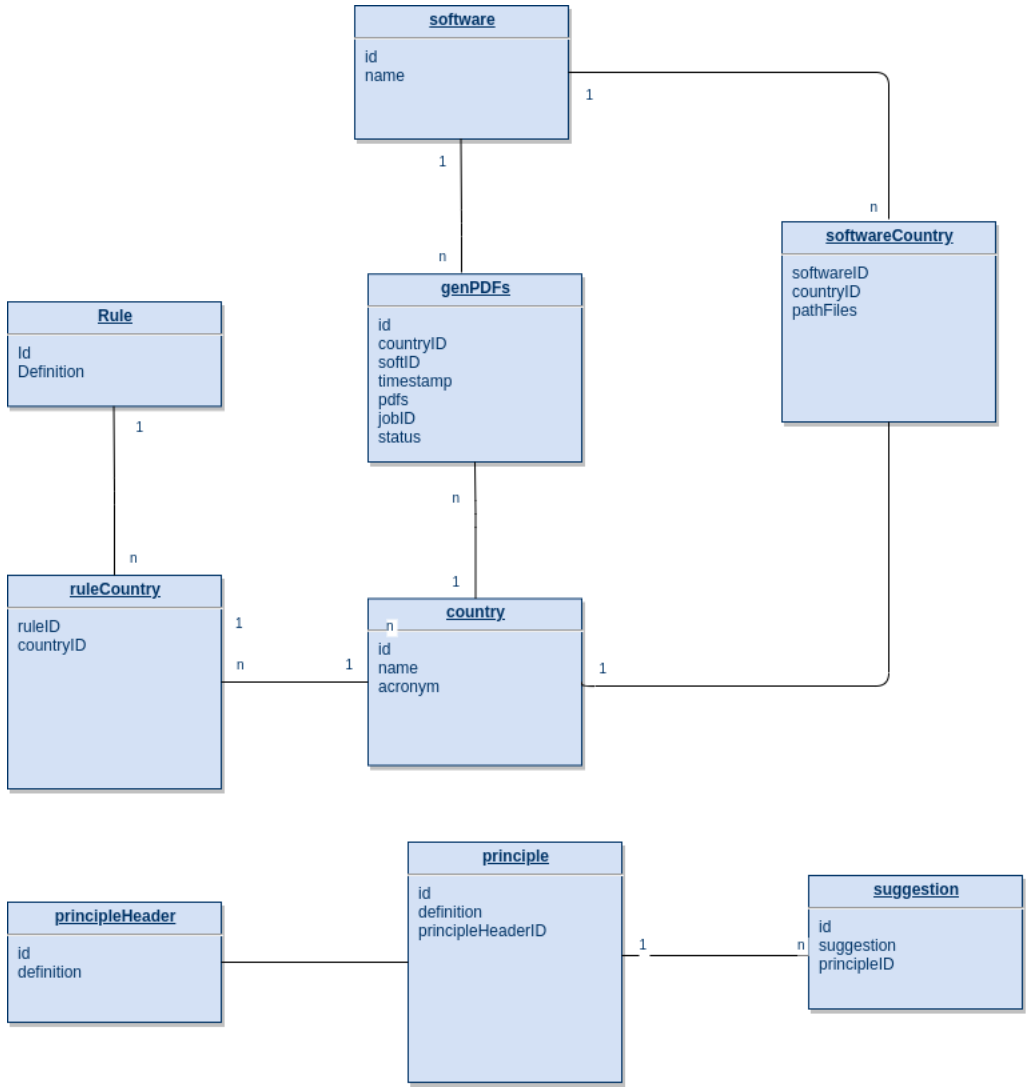


Figure 4.3: Database schema

composition of those tables is based on the articles studied previously and make use of the seven principles which we defined. On top of that the table *suggestion* exists because if for some reason the application does not complies with a certain definition, on the final report will appear that suggestion. This information already comes pre-defined in the database. Also, the information about the principle definitions and suggestions was extracted manually while studying the GDPR regulation.

As discussed before, there might be a case where a country besides applying the GDPR regulation also has to comply with other rules specific to that country. To support this requirement the tables *rule* and *country* were included where any a country can have multiple rules and also the same rule to be used by multiple countries. Also, the *country* table connects to the *software* one, so when the user on the frontend application chooses a country it automatically filters the list of software related to it and also loads the rules for that country.

Finally, the table *genPDFs* was created to store all the reports and is directly related with the table *country* and *software* since every report is created based on that information.

4.4 PADRES Workflow

The following Sequence UML Diagram describes a common interaction between the user and the application. The goal is to give an overview of the process and also to understand the following sub sections easily.

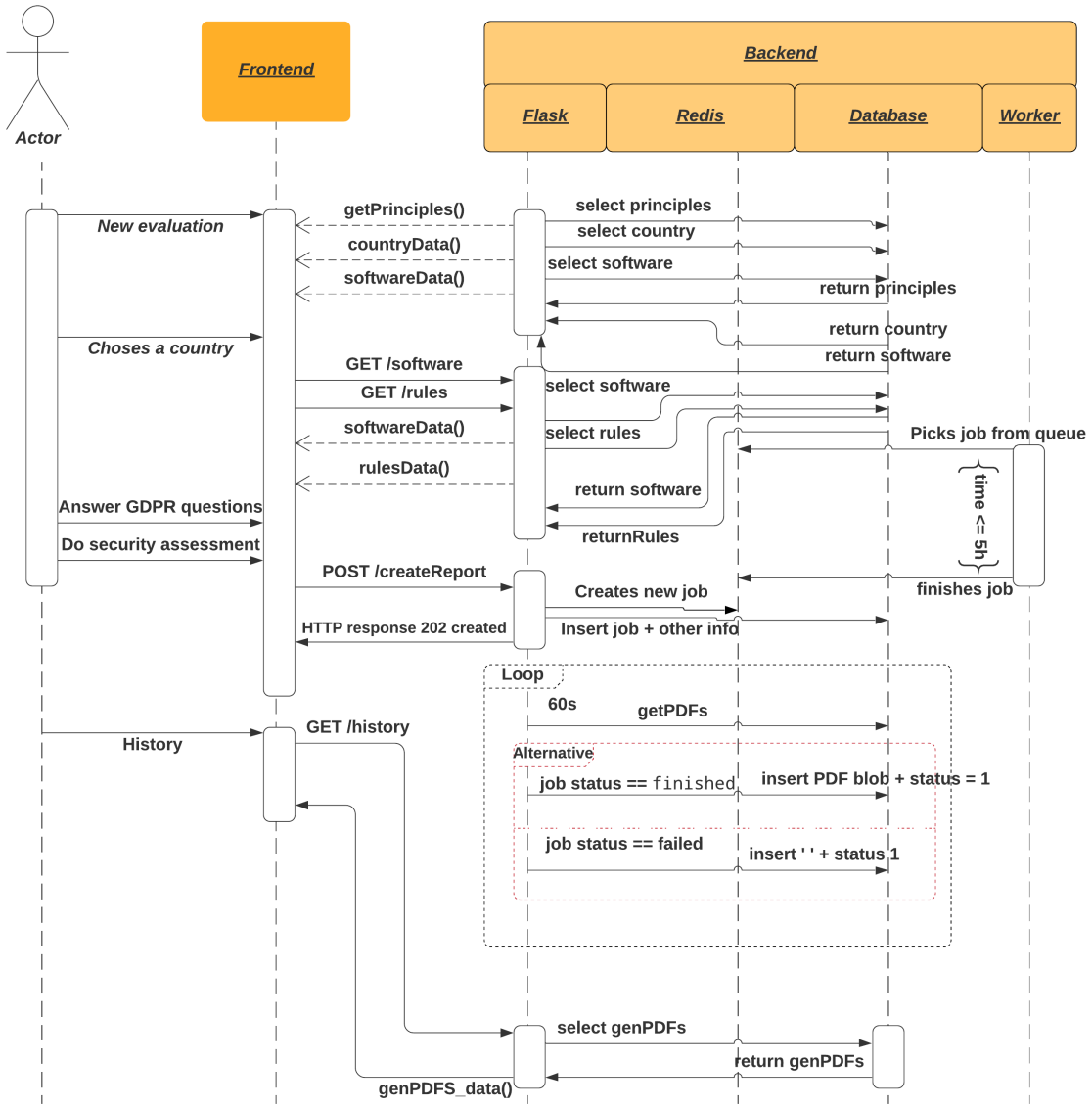


Figure 4.4: UML sequence diagram representing a common application interaction

This diagram describes the application common use, where the actor chooses to do a new evaluation until it proceeds to submit it to the backend where the final report is created. A detailed explanation of each step is given on the following sections

4.5 Configuration

In order to streamline the solution's distribution and execution, Docker was used to provide those advantages. Docker is a platform that enables a easier way to build, deploy and run applications using containers. This is possible due to its level of abstraction, by giving to each

container their own ”sandbox”, packaging together code and dependencies. This allows a faster deployment and also increases the reliability of a solution, since it does not depend on the external environment configuration. So, each container must have their own definition in a file called *Dockerfile*, where is expressed what needs to be done to run properly.
Docker itself can only handle single containers, so one of the problems comes when there are several containers to be managed. To solve that is possible to use *Compose*, allowing to define and run multi-container Docker applications. This is achieved through a document, called *docker-compose.yaml*, that merges together the definition from each *Dockerfile* and also giving the possibility to have their own network setting and volumes. Based on that, the PADRES follows that methodology and the diagram 4.5 describes that.

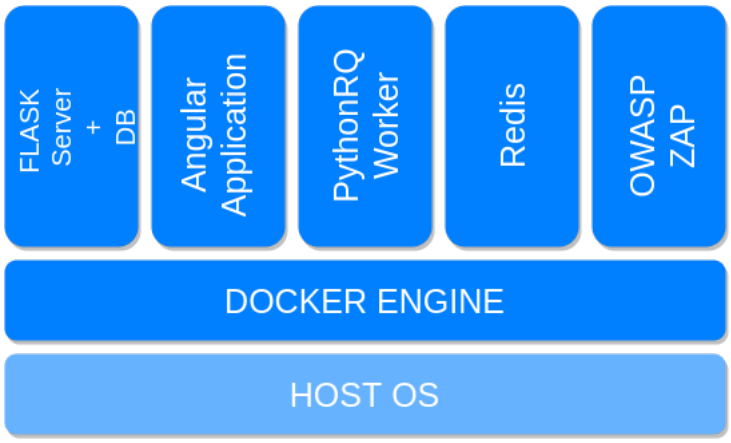


Figure 4.5: Docker diagram

From image 4.5, is possible to see there were defined five containers, deployed together using the *Compose*, that were already explained on the sections 4.6 and 4.7. Basically it encompass all the components that were used in a single one.
Also using this structure is possible to add more security assessment tools to the current project. Currently this process involves to study how the tool behaves and then it can be called from the *Flask server*. So it can only be added manually. The best tools to use are the ones that provide an API, but most of them are not free and that is why ZAP was chosen. Also the Wapiti, that does not provide an API so their output options need to be evaluated in order to choose the best one, which adds more complexity to the system.
To run the solution, Docker and Docker-compose must be installed. Once done the user must use the following command to build and start up the solution and in the future it only has to use the command without the *-build* option.

```
$ docker-compose up --build
```

4.6 Front-End

The front end interface aims to offer the final user a view of the principles and the corresponding regulation points. Also, based on the analysis done, it is also provided a section to check the history reports. Since this is only for proof of concept, this interface was kept simple and easy to navigate.
When the user chooses to do a new review of their compliance, he must choose which software and in which country the review must be based on. This is the first step, the choice of software

needs to be done, because one company can have multiple softwares and even if they are used together, as studied in other cases, it is simpler to analyze when they are spited. The need to choose the country, comes to support the ones, that have others laws besides GDPR that needs to be fulfilled. This selection, is done through a drop down select using the Angular Material select, which is filled by doing a HTTP request to the backend, which answers with all the softwares and countries. Based on the database schema is possible to visualize that the relation between software and country is many to many, which allows, in the interface, to filter the countries when a the user chooses a software and filter the softwares when the user chooses a country.
After the selection is done, all the principles are loaded, including the one specific for each country. To achieve this, it was used the Observables interface from the RxJS library, providing a way to handle asynchronous operations. This type of operations are used to send data between components or, as in this case, to do HTTP requests, allowing the program to proceed, without having to wait for data to arrive.
The following piece of code describes the process of requesting data in Angular through the the HTTP Client, that is a simplified client HTTP API, and using the Observables interface to handle the response.

```
getPrinciples(phID: number): Observable<any>{
    return this.http.get(baseURL + 'principles/' + phID).pipe(
        catchError(err =>  throwError(err))
    )
}
```

This function receives a number by parameter, *phID*, which represents the principle identifier, from which we want to have the linked points and returns an Observable, *Observable<any>*. Then passes it to the *get()* function the Universal Resource Identifier (URI).
When requesting data from another server, there is the need to prepare for something to go wrong. That's why the response is piped through the *catchError()* function, intercepting a request that failed, and passing the error to a possible handler or, as in this case, throws it, so it can be handled later.
Since this function is defined inside a service, this must be injected in the desired component, more precisely in the class constructor. The following code snippet shows this process.

```
constructor(private principleService: PrincipleService,
private fb: FormBuilder, private router: Router) {
    this.createForm();
}
```

The constructor is the first method to be called when initializing a component, which in this case is defined by the *principleService* type of *PrincipleService*, where is the function *getPrinciples()* is located.
The next snippet describes the mechanism of subscribing an observable. By iterating over the results being sent, we can do everything with them. In this case, they are stored and then displayed to the user.

```
this.principleService.getPrinciples(parseInt(principle.id)).subscribe(
    (data: Principle[]) => {
        data.forEach((d: any, index1) => {
            ...
        });
    },
    error1 => {
        console.log(error1.err.message);
    }
);
```

Considering that there was some error while doing the request, is possible to do some logic with that, like retry the request or alert the user, but in this case it is only displayed a message in the browser console.
The remaining data needed, is obtained following the process that was being explained, as well as the request to send data to the server.
Finally, when the user chooses to submit the data, it is shown a pop up asking if him wants to do the security assessment. If the answer is positive, then he has to insert the required data and after he is redirected to the history page, where is possible to see the previous analyses, as well as, download the report generated in PDF format, when available.

4.7 Back-End

To develop the backend API,Python was used as programming language and Flask as the web framework. The reason to choose those two, is mainly due the fact that they are easy to get started, lightweight and both have a great documentation, ensuring at the same time the ability to scale up to more complex applications. Also will be approach how the GDPR questions were obtained and how they stored and accessed.
Other libraries were used and will be detailed on the following sub sections.

4.7.1 Python and Packages

Starting with Flask, which was relatively easy to set up by following the documentation, and after that endpoints routes were added so it is possible to interact with data. For example to get the regulation point for each principle it was used the following code.

```
@app.route('/principles/<phID>', methods=['GET'])
def principles(phID):
    dbCon = None
    try:
        dbCon = conDB.newCon()
        data = conDB.getPrinciples(dbCon, phID)
        response = app.response_class(
            response=jsonParser.principlesJSON(data.fetchall()),
            status=200,
            mimetype='application/json'
```

```
        )
        return response
    except Exception as e:
        abort(500, {'message': e})
    finally:
        if dbCon is not None:
            try:
                dbCon.close()
                app.logger.info("dbcon closed {}".format(dbCon))
            except Exception as e:
                app.logger.error("Error closing con {}".format(e))
```

This endpoint is available through the URL */principles/<phID>* and is marked with *GET*, assuring that it only answers to HTTP GET requests.
Then, it is used an auxiliary library with a set of functions the interact with the database, as well as one, to convert the data obtained to JavaScript Object Notation (JSON).
Finally, is was created a *response_class* object which will be returned, containing the JSON with the HTTP code, that in this case is 200 since it was a successful request and also sending the type of data returned.
If for some reason there is an error, it will fall in the *except* section returning a 500 HTTP code that correspond to a Internal Server Error with the error traceback.
More endpoints like this were made to answer the application demands, with special attention to the endpoint responsible to handle the data sent from the frontend through a *POST* request. This one, handles the user input in order to create the final report in PDF format.
Dissecting this function that encompass the GDPR compliance analysis and the security assessment goal, it starts by looking to the regulations points which had a negative answer and fetch from the database the suggestions to be applied in order to improve the compliance. Once it is done, an HTML file is created with the information, to smooth the user reading. Then, depending on the user choice, is performed the security analysis. Since this one takes some time until it completes, depending on several factors that will be discussed later, was created using the *PythonRQ* library a job, which will be handled by a worker, so it does not block other requests. When the job is created it returns a job identifier, giving then the ability to check in the future its status and also to execute other operations. This setup is defined with docker and will be explained on section 4.5. The worker entity mentioned above, also defined in the textitPythonRQ , runs on a different container and is backed by *Redis*, a in memory database used for cache services purposes.
Finally, is created an entry in the table *genPDFs* with the previous information and with status of 0, meaning that the final report is not ready yet.
In order to keep track on the jobs in queue, was developed a function that runs on the background every 60 seconds. When the job return the finished status, it updates the entry with that job identifier, becoming now with a status of 1 and also storing the final PDF file inside the database.
Inside the job, which is defined in this case by the function *doAllScans()*, is then executed the security analysis. This consists in putting together *NMAP*, *OWASP ZAP*, *Wapiti* and a cookie scanner. In the end each one produces a report and the final result is a PDF with all the information generated.

4.7.2 GDPR questions

Throughout the dissertation, the GDPR subject has been studied through all the articles and all regulations web pages already cited. In a certain way, it was summarized and written on the previous sections. The goal of this study was to obtain the knowledge in order to summarize the regulation into several points, so companies it limited resources could have a first approach on how to be in compliance.

Besides those citation, is important to refer the following ”letter” [106], that fits exactly into this situation. This ”letter” was written to describe the worst case scenario possible that a company can come across. The author even says, that it can be used to perform table-top exercises, a technique to prepare for the eventuality of a disaster scenario. In this case it aims to verify how a company would react to the possible case of someone with knowledge in law and technologies to support data management, to send a list asking and demanding his rights. This ”letter” was then used to help writing the points, making them more precise.

Based on that, was concluded that the most appropriated framework for the regulation, was to split it into principles, as seen on table 2.2. Then inside each one, a list with the most important points that need to be followed. On figure 4.3 is possible to see that the database was designed with that in mind.

The points for each principle were then extracted manually, keeping each one simple and concise enough, to not cause any confusion in whoever reads it. For example for the the first principle were made the points *”Does the consent inform the Individuals about the processing objectives”* or *Does your application provide any information regarding the Individual’s rights* and so on. Those points and principles are already available on the database, being be access through the application the best way to access them or directly through the API using the endpoint */principles/<phID>*.

Finally to develop the suggestions it was also used the same sources of information, but this time to convert them into not rules but approaches possible to be implemented. Even though some are more technical than others, the goal is to improve the level of compliance. For the current project were written suggestion for the rules that can be more difficult to understand and implement. For example related to the rule *Do you have any mechanisms to pseudonymize data?* if the user answers no, then on the report is shown the suggestion *”Its a data management technique, where the data controller swaps the individual’s direct identifiers, such as email or phone number, with a pseudonym. Then the data processor can process the data without exposing the sensitive data. Then when data goes back to the data controller, he can rebuild the original data through re-identification techniques”*

4.7.3 External tools used and cookie scanner

The solution provided makes use of external tools that are integrated inside PADRES. This subsection describes how this is achieved.

Starting with *NMAP* it was chosen due to the fact that not only looks for open ports but also can produce network exploration among other functionalities. In this specific case will be used only against a single hosts, provided by the user on the frontend application. As said on their website [107], the unappropriated use, can lead to ”ISP account cancellation or even civil and criminal charges”, so the final user must take this into account when using it.

In order to use it in this solution, was used the library *python-nmap* [108], which makes the automation and the output manipulation easier to whoever uses it. To use this library, firstly, is

instantiated the nmap object and then is possible to call the scan function that take as arguments the target and also additional arguments which will compose a *NMAP* command. In this situation the scan function called looks as follows:

```
nmScan.scan(hosts=target,
            arguments='-A --script=nmap-vulners/vulners.nse')
```

Dissecting the function above, when is called without the arguments parameter it produces the following *NMAP* call:

```
$ nmap -oX - -sV [target]
```

However when called with the additional arguments simulates the *NMAP* command

```
$ nmap -oX - -A --script=nmap-vulners/vulners.nse [target]
```

This was possible to check due to the function *nmScan.command_line()* , that returns the original *NMAP* command, allowing to full understand what is going to be analyzed.
Dissecting the command, it adds the option -oX, that basicly gives the output in Extensible Markup Language (XML), the - option so it does not write the interactive information to the Stdout but only to the XML. The -A option, acts as a shortcut for the options *-sC*, *-sV*, *-O*, *-traceroute*, which enables script scanning, version detection, OS detection and traceroute, respectively. The script scanning inside *NMAP* using the Nmap Scripting Engine (NSE), is classified by the authors ”one of the most powerful and flexible features”, allowing the users to write scripts to be run inside *NMAP*. Since one of the options chosen were the -A, that uses -sC, it tells the NSE to use the default set of scripts, which was put together by *NMAP* team as a collection with the most popular ones, for example the *mysql-info* to check for information about a MySQL server or the *http-git* to find if there is a *.git* folder inside the website document root.
The last option –script, runs the vulners.nse script. This option is the same as the one explained before but in this case is set to specifically run the desired script. This one, made available by the Vulners Team in their Github, contains a set of CVEs giving more information on vulnerabilities to the final user
Finally, the function *nmScan.get_nmap_last_output()* is called to get the output in XML format so it can be converted to HTML using the *xsltproc* tool after.
The second tool, OWASP ZAP is focused on search for common web vulnerabilities. This tool fits in the scope of penetration testing and acts as a MITM agent by standing between the user and the application being tested. Even though this tools offers a good desktop application, in this case will be used the ZAP API, which allows for automation and also integration with python, offering at the same time almost all the feature found on the desktop app.
This tool can be used by simply download and execute it, or, as used in this solution, using their docker image and since the solution presented here is already supported by docker and the acZAP API runs as server, makes sense to run it in a different container. Besides that, the acZAP API provides a Software Development Kit (SDK) for python, which was used to access to API instead of calling the endpoints directly.
To start using it, firstly is created a *ZAPv2* instance and passing as arguments the proxy address, that in this case looks as follows

```
zap = ZAPv2(proxies={'http': 'http://172.19.0.3:8090',
                     'https': 'http://172.19.0.3:8090'})
```

Then in order to test the desired application for vulnerabilities is recommended to run a crawler that will explore it entirely and at the same time builds a ”map” of all the end points reached by

it. ZAP provides two types of crawlers, the spider scan, that finds HTML resources and the Ajax Spider, used for applications based on Ajax calls and JavaScript, which nowadays are heavily used. In this solution both scanners were used in order to obtain a broader field of possible tests. The following piece of code shows how to start the spider and Ajax scanners and waits for them to finish before proceeding to the next instructions

```
# spider scan
scanID = zap.spider.scan(target)
while int(zap.spider.status(scanID)) < 100:
    time.sleep(2)

# ajax scan
zap.ajaxSpider.set_option_max_duration(5)
print(zap.ajaxSpider.option_max_duration)
scanIDajax = zap.ajaxSpider.scan(target)
while zap.ajaxSpider.status == 'running':
    print('Ajax Spider status' + zap.ajaxSpider.status)
    time.sleep(2)
```

Even though, they are both scanners, they act differently. While the spider scan has a status allowing to know when it ends, the Ajax one, does not give that, so instead of waiting indefinitely, was set the option *set_option_max_duration(5)*, allowing it to run only during five minutes.
Is important to note that while both crawlers are running, at the same time and by default, the resources being found, are tested against a passive scan. This is one of the attacking methods of this tool, and tries to find weaknesses on HTTP calls or to look for anti Cross-site request forgery (CSRF) cookies. For example this two possible ways of attacking the application fit inside the Security Misconfiguration and the Broken Access Control respectively of the OWASP TOP 10 vulnerabilities. To obtain a list of the attacks that will be performed, the final user should access the ZAP API through the browser, specifically on the URL *http://localhost:8090/JSON/pscan/view/scanners/?*. Also following the same method on the endpoint *http://localhost:8090/UI/pscan* is possible to find options to disable, enable all scanners or even only enable the desired ones among other options. In the presented solution, by default all the scanners are enabled.
Another method of attack present in this tool is the active scan. This one instead of just listening, it actually attacks the application to find potential vulnerabilities. Also like the other method is possible to get a list of all the scanners and perform another options to get the best of this tool.
Since the *ascan* attack provides multiples scans and option to disable and enable them based on a given policy, it was set to start it the default one. Those attack policies, depending on the chosen one, besides activating the attack rules also define the Threshold and Strength. The first parameter sets the minimum level for ZAP report potential vulnerabilities and has three levels. Depending on the chosen level, it can lead to the increase number of false positives or even worse more false negatives. The strength parameter defines the number of attacks being done, that increases the time for a scan to finish.
In addition to those option there is the attack mode, that can be one of the following options. The safe one not allowing dangerous operations, the protected, which only attacks URLs in the scope, the standard allows to do everything and finally the attack mode attacking the new nodes when they are added to the scope. By default ZAP sets the mode to protected.
These combinations of options will be explored on the chapter 5 in order to analyze the one

that produces better results.
Another way to interact with the scanners and their options is through the OWASP ZAP API SDK as can be seen on the following piece of code:

```
#passive scan
zap.pscan.disable_all_scanners()

#active scan
zap.ascan.disable_all_scanners()
zap.ascan.enable_scanners(40018,40012)
```

When all the scans come to an end, the function *zap.core.htmlreport()* is called, generating a report with all the alerts raised when the web application was under attack.
The third, *Wapiti*, which is also focused on web vulnerabilities, does not offer the same level of features and options as the ones made available by OWASP ZAP. It was chosen, to give to the presented solution redundancy and possibly find vulnerabilities that were not found using OWASP ZAP and vice versa. When this tool starts a scan, firstly it also does a scan to find all the available resources in order to have a ”map” of the application and only after that the attack begins.
Since this tool can only be access by command line, was used the *subprocess* library from python to call it enabling to pass a command to it and obtain to output through the *subprocess.PIPE*.
The command structure used looks as follows:

```
command = ”wapiti –u” + target + ” –m sql,blindsql,xss,permanentxss,
htaccess ––flush–session –f html  –o ./pdfs”
```

Dissecting it, the option *-u* defines the target, the *-m* defines the attack modules to be used, which were chosen to compare with the ones from OWASP ZAP and adding on that, also the verification of *.htaccess*, a file used to configure web servers running the Apache Web Server, searching for bad configurations that is one of the vulnerabilities found on OWASP top 10. The option *-flush-session*, tells *Wapiti* to clean previous results related to that target and the options *-f* and *-o* to set the output to be of type HTML and to put it in the folder *pdfs*.
Finally the cookie scanner. As studied before, the GDPR regulation says that is possible to identify a person through cookies, due to the fact that this small text files can store not only crucial information for a web app to run but also to store information about the user himself. In that in mind is possible to state that consent must be given before using cookies, except the ones necessary for the application life cycle to run normally. Besides that the ePrivacy Directive also states the same as the GDPR about cookies, saying that a consent must be given and also the user must receive a clear and comprehensive statement about what data is being collected and its purpose.
Based on that, was written a scanner that simulates the first user access to a web application to gather the cookies stored. Then it proceeds to search in the source code for any button associated with accepting a consent. If found, simulates the click action and again gathers the cookies stored. This scanner was written in Python using the *Selenium* framework to interact with the browser.
Due the various number of third-party cookie managers available and being that each one of them has different implementations, means that it may not be possible to do this procedure in all the applications. So, to overcome that issue, it will only search for cookie managers based

on the *Cookiebot* solution, *Quantcas*, *OneTrust* and *CookieScript*, which are among the most used softwares for that purpose.
Is expected that on the second harvest the number of cookies to be higher that in the the first one.
Once all the scans mentioned above are done, is returned all the previous HTML files, that will be converted into one final PDF, using the library *pdfkit*.
As said before, all of this tools are called inside one function. Once it returns, also closes the job associated with it, passing the responsibility now to the listener. This has as main function to update the database, that consequently makes the PDF available on the frontend, through the API so that the final user can read it. Is important to have in mind that those security scanners may take several hours to finish and consequently the final PDF will also be delayed. In the end of the PDF the user can see the classification based on the questions answered positively out of all the questions available, so it can have a metric to compare with future or previous assessments.

Chapter 5

Tests and Results

5.1 Introduction

This chapter presents the tests made in order to validate the solution and also a discussion of these results in order to determine what can be improved and also to analyse the changes possible to be made for each scenario. The focus will be on the GNSS Products Gateway and GNSS Data Gateway, since the framework of this dissertation is the EPOS project.

5.2 EPOS GNSS inspection

The tests made for EPOS were relatively straightforward and done by the author of this dissertation, since he was also member of this project and was involved directly in its development. Taking into account the EPOS GNSS environment, firstly the landing page will be evaluated and then to understand how deeply the security assessment tools can go, each one of the sub web pages will be evaluated and in the end compared between them. Also, since some of the pages uses personal data independently, the GDPR question will also be answered.
Currently EPOS softwares are listed in the following link *https://glass.epos.ubi.pt:8080 /Glass-Framework/*.
Running PADRES on the URL mention above, the GDPR questions are left unanswered because it does not collect any personal data so the GDPR regulation does not apply. When submitting the form, were checked the boxes to run NMAP, OWASP ZAP, Wapiti and a cookie scanner against this URL.
Analyzing the full report generated after all scans conclude, comes the information related to the cookies found, that were none. By looking on the devtools available in a browser, is possible to say that the information given is correct.
After that comes the NMAP report. It was found that the URL provided, has the IP address of 193.136.66.9. Also found that 948 ports were closed by sending a reset response. Besides that was found 10 open ports, identifying the service running on each one and since it was given to NMAP the script option, it also searched for vulnerabilities on each service. From the image 5.1 is possible to see that the port 21 has the FTP service running, representing some issues since the authentications uses plain-text passwords, so any MITM attack can see the information being exchanged and then compromise the system. So to overcome this issue is recommended to use SFTP. The port 22 is also open accepting ssh connections, using the OpenSSH service, on version 5.3 and using protocol 2.0, which offering authentication using SHA-2 hash algorithms instead of the SHA-1 used on the previous version or authentication using FIDO/U2F [13]. Besides that, is included a list of the CVE, which is list of entries containing publicly known cybersecurity vulnerabilities, for that service using the version 5.3. Based on that is possible to follow the link available or by searching it on any search engine to get a vulnerability description. Next, comes the port 80, that follows the same methodology as port 21 described above, but now it refers to the HTTP server running apache server. Following the NMAP report, it also, on port

Port		State (toggle closed [0] | filtered [0])	Service	Reason	Product	Version	Extra info
21	tcp	open	ftp	syn-ack	vsftpd	2.2.2	
22	tcp	open	ssh	syn-ack	OpenSSH	5.3	protocol 2.0
	vulners	cpe:/a:openbsd:openssh:5.3: CVE-2010-4478 7.5 https://vulners.com/cve/CVE-2010-4478 CVE-2017-15906 5.0 https://vulners.com/cve/CVE-2017-15906 CVE-2016-10708 5.0 https://vulners.com/cve/CVE-2016-10708 CVE-2010-5107 5.0 https://vulners.com/cve/CVE-2010-5107 CVE-2016-0777 4.0 https://vulners.com/cve/CVE-2016-0777 CVE-2010-4755 4.0 https://vulners.com/cve/CVE-2010-4755 CVE-2012-0814 3.5 https://vulners.com/cve/CVE-2012-0814 CVE-2011-5000 3.5 https://vulners.com/cve/CVE-2011-5000 CVE-2011-4327 2.1 https://vulners.com/cve/CVE-2011-4327					
80	tcp	open	http	syn-ack	Apache httpd	2.2.15	(CentOS)
	http-server-header	Apache/2.2.15 (CentOS)					
	vulners	cpe:/a:apache:http_server:2.2.15: CVE-2011-3192 7.8 https://vulners.com/cve/CVE-2011-3192 CVE-2017-7679 7.5 https://vulners.com/cve/CVE-2017-7679 CVE-2017-7668 7.5 https://vulners.com/cve/CVE-2017-7668 CVE-2017-3169 7.5 https://vulners.com/cve/CVE-2017-3169 CVE-2017-3167 7.5 https://vulners.com/cve/CVE-2017-3167 CVE-2013-2249 7.5 https://vulners.com/cve/CVE-2013-2249 CVE-2012-0883 6.9 https://vulners.com/cve/CVE-2012-0883 CVE-2018-1312 6.8 https://vulners.com/cve/CVE-2018-1312 CVE-2017-12171 6.4 https://vulners.com/cve/CVE-2017-12171					

Figure 5.1: NMAP result for *https://glass.epos.ubi.pt:8080/GlassFramework/*

443 it detected a GlassFish installation, using the SSL service, probably because the sites hosted on Glassfish uses SSL/TLS certificates to protect the data being exchanged between the server and the user. Additionally port 8080 has the same installation of the service on port 443, but hosts the webpage being attacked. Also has found the port 8081 and 8082 to be open, using the SSL service and also hosts web pages being used for the EPOS project.
Now proceeding to the ZAP report, the scan is resumed on the image 5.2.

Summary of Alerts

Risk Level	Number of Alerts
High	0
Medium	3
Low	8
Informational	3

Figure 5.2: ZAP scan result for *https://glass.epos.ubi.pt:8080/GlassFramework/*

The scan as said on the previous chapter uses the default policy, which sets the attack strength and threshold to medium, resulting on rising no alerts for high risk vulnerabilities, but rising three for medium risk. The report then minutely details where the vulnerability was found, how many instances of that vulnerability were found and also a solution to fix it. Also gives a Common Weaknesses Enumaration (CWE) and Web Application Security Consortium (WASC) identifiers to search for more information about the risk. For example the first vulnerability found was the *X-Frame-Options Header Not Set*. This one fits inside the Security Misconfiguration of OWASP top 10 and is a security header that comes inside the HTTP requests, defining if a page can be rendered inside HTML elements such as <*iframe*>. For that header there are two options available. The **DENY** and *SAMEORIGIN*. The first blocks the page from being rendered anywhere and the second only allows the page to be rendered inside a page that has the same origin [109]. More information can be found searching the CWE 16 and WASC 15.
Besides the vulnerabilities found with medium risk, also was found eight with low risk and three

informational. The low risk ones, encompass vulnerabilities such as*X-Content-Type-Options Header Missing* or *Cross-Domain JavaScript Source File Inclusion* referring to the use of external libraries, among other mostly related to missing HTTP response headers. Those vulnerabilities are not affected by the attack strength and threshold since it only affect eh active scans and the vulnerabilities related with cookies are done by the passive scan.
Finally the wapiti report did not report any vulnerability.
Based on the report was possible to analyze that the web application referred on this page, were not scanned. This because the ZAP and wapiti, were set to only crawl for results inside the scope, which is the base URL. This can also be proved by searching on ZAP API specifically on endpoint *http://localhost:8090/..* and in wapiti by setting the scope option to domain. If other option was set, it could lead to unwanted scans and future problems.
Based on that, was scanned another web application, also inside EPOS, called *GNSS Products Portal*, available on the URL *https://gnssproducts.epos.ubi.pt/*, with the same options as the previous one.
This one collects personal data due to its authentication mechanism, so the GDPR questions were answered giving the final result of 8 not in compliance points from a total of 31 points. Is important to enhance that those questions were answered by José Manteigueiro from de EPOS project, who is the authentication system developer.
To those points that the system do not comply with, some suggestions were given. For example, the point *Have you performed any audit to map data flows?*, suggests some measures as can be seen on the image 5.3

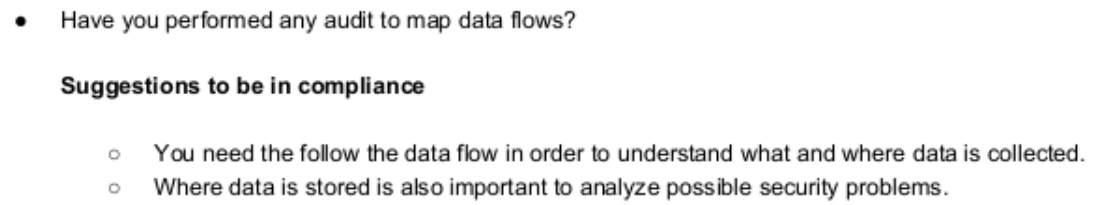


Figure 5.3: GDPR suggestion example

The following 7 points also follow this structure, even though some of the points do not have suggestions since the questions is itself explicit enough to know what to change.
Also is reported the use of two cookies when a user access the web application, the *laravel_session* and *XSRF-TOKEN*. This first one is used to identify a session instance for a user and the second to prevent unauthorized events to be performed on behalf of an authenticated user using is authentication cookies. Then it was detected the use of a cookie manager to manage the consent and as explained on the previous chapter, it performed the click to accept the consent and was obtained the cookie *CookieConsent* with a stamp value. Together with these information, comes the information if a certain cookie is *HttpOnly* or not. This field is defined in the response header and helps to mitigate the risk of client side scripting. In this situation only the cookie *laravel_session* as the *HttpOnly* set to true.
From the NMAP scan is possible to conclude that this application is located on the same IP as the previous application so the result was the same.
Regarding the ZAP scan 5.4, as can be seen of the following image, reported two high risk vulnerabilities. The first vulnerability with that risk has the name *Path Traversal*, meaning that the attack was able the access files and directories that should not be accessible at all. Normally

Summary of Alerts

Risk Level	Number of Alerts
High	2
Medium	5
Low	16
Informational	5

Figure 5.4: ZAP scan result for *https://gnssproducts.epos.ubi.pt/*

the website files are located inside the web root folder, */var/www/html*, and with this type of attack, the attacker is able to access the file outside, for example /etc/passwd.
In this particular situation, as can be seen of image 5.5, ZAP explored the URL *https://gnssproducts.epos.ubi.pt/auth/try/yubikey* by sending a POST request with the parameter *confirm* with value *c:/* resulting in accessing the folder */etc*. The same was done for the folloowing URL but this time with the parameter *remember* and with value / resulting in the same folder.

URL	https://gnssproducts.epos.ubi.pt/auth/try/yubikey
Method	POST
Parameter	confirm
Attack	c:/
Evidence	etc
URL	https://gnssproducts.epos.ubi.pt/login
Method	POST
Parameter	remember
Attack	/
Evidence	etc

Figure 5.5: Path Traversal attack on *https://gnssproducts.epos.ubi.pt/*

In order to ascertain the results, we tried to reproduce the request using the *CURL* command and also in the browser but the results were an error page.
This attack was done 33 times by ZAP all with the same result, *etc*, so is possible to assert that is was a false positive. However, the attack gave some solution to prevent this type of attacks, that can be the implementation of ”accept known good” input strategies, running the code with the lowest privileges possible or running the code in containers where is provided abstraction with the host.
The other high risk vulnerability found, named *Remote OS Command Injection* and as the name suggests, occurs when a vulnerable application allows the execution of commands on the host OS and are often related with insufficient input validation.
In this situation, was used the same methodology used above. ZAP tried to inject commands using POST requests, using values such as *ZAP;sleep 15;*, that simply tries to delay the program execution.
To verify the accuracy of the results, was again executed the same requests, but again landing on the same error page, meaning that again was a false positive attack. These vulnerabilities were found by the active scan and are affected by the attack strength and threshold given.

Yet, is important to understand the suggestions to avoid this type of vulnerabilities. For example use library calls instead of external processes to perform some functionality or sanitizing the inputs.
The vulnerabilities found for the other risks, were all based on misconfiguration of the header requests and were already approach above, highlighting a new one found that is *Absence of Anti-CSRF Tokens*. This one was found when submitting a request. So based on the cookie scanner mention before was found the cookie *XSRF-TOKEN* and when analyzing the request in browser the development tools that cookie is set when doing the request.
Relatively to the wapiti scan it did not report any vulnerability.
Since all these security tests were done using the scans default configurations, to possibly obtain different results, they were changed. The GDPR and cookie components remain the same, while the in ZAP the attack Strength was set to High and the Threshold also to High. This configuration will produce more attacks to each endpoint, however with that threshold, the number of false positives is expected to be lower. Regarding the wapiti the attack level was set to 2, that increases the attacks payloads even if there is no parameter present in each endpoint found. Based on the tool documentation, this option is not recommended, because the ratio between success and attacks was not good, since the previous configuration did not report any vulnerability, which can be a good point. Also was set the option *-scan-force* to aggressive, that increases the requests sent based on the input parameters an URL or form may have.
With that configuration, for the URL *https://gnssproducts.epos.ubi.pt/*, the result is summarized on the image 5.6.

Summary of Alerts

Risk Level	Number of Alerts
High	2
Medium	3
Low	15
Informational	4

Figure 5.6: ZAP scan result *https://gnssproducts.epos.ubi.pt/*

The result, comparing with the previous attack, is the same for the high risk vulnerabilities and the vulnerabilities reported are also the same. If in the previous attack was possible to say that the results were false positives with a threshold of medium. Now with a threshold of high, was expected that possibly those vulnerabilities would not be reported. Since the attack the set also to high, a bigger number of attacks were also done, being reflected in the number of instances created, which were 37 in this report comparing with 33 from the previous one.
The medium risk vulnerabilities this time were lower, being again the reported as *X-Frame-Options Header Not Set*. This decrease is not affected by the configuration given above, so it can happen that the passive scan have had a different behaviour in this scan.
This time the GDPR scan was not done since it is the same from the previous report and the cookie scanner reported the same cookies. Regarding the Wapiti scan it reported again no vulnerabilities.
Also with the previous configuration was tested the application *https://glass.epos.ubi.pt:8080/*,

that when comparing with the previous report from the same URL, did not report any high risk vulnerability and the number of medium and low risk vulnerabilities were the same.

5.3 C4G intranet

The following tests were done on the C4G intranet, that is a Collaboratory providing services related with geo sciences subjects. The intranet gives the users the possibility to login and register, and from there interact with the C4G platform. So the URL in study is *https://intranet.c4g-pt.eu/*
Since is also provides a register, then is possible that personal data is being collected, so it is a good platform to test the performance of our solution.
Following the approach done in the previous tests, firstly the will be used a default configuration for the external tools and then they were changed in order to see the differences.
The GDPR questions were answered by one the developers and maintainers of the platform, Luis, resulting in a non compliance of 15 point out of 31. For example one of the points with no compliance is the one shown in the following image 5.7. From that is possible to say the possibly that, even though not using pseudonimization is not a major issue, with the suggestion the user is more aware of the pseudonimization purpose and if necessary implement it. The next point in that image regarding encryption was also answered as not in compliance, that represents a major issue if true. Possibly the user answered wrongly this point because for example when accessing their intranet the connection is done through the HTTPS. In this point there was no suggestion shown due to being a topic that everyone in this field knows, even though, is always recommended the search for best practices. The image on section A.2 corroborates the affirmation done.

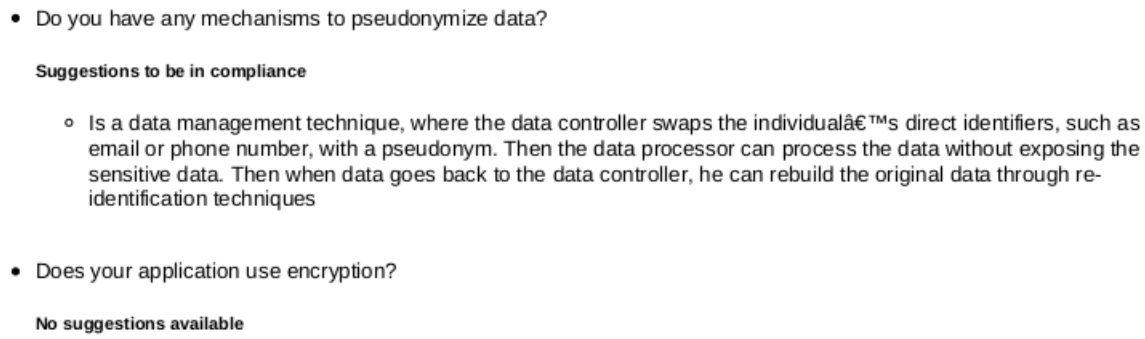
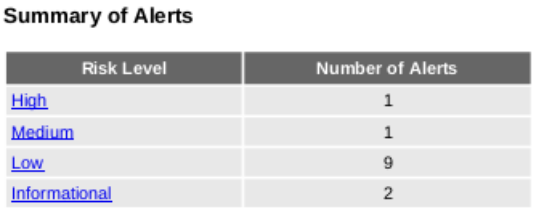


Figure 5.7: C4G GDPR point not in compliance

The following item from the report is the cookie scanner, showing that only two cookies are being saved. One related with the PHP session and the second, the XSRF-TOKEN, to protect against CSRF attacks and was already explained before.
Now looking to the NMAP report is says that the URL mention above is hosted on the server with the IP address of *193.136.66.9*. This one is the same as the EPOS GNSS from the section 5.2, so the vulnerabilities found, are the same as in this one.
Moving to the ZAP using the attack strength and threshold as medium it reported only one vulnerability classified as high, one as medium and nine as low as can be seen on the image 5.8. The one reported has high refers to the path traversal vulnerability, which was already identified and described on the previous test. In this case was again tried to replicate the attack in order



Summary of Alerts

Risk Level	Number of Alerts
High	1
Medium	1
Low	9
Informational	2

Figure 5.8: ZAP report for C4G

to see if it is a false positive or a True positive. So, trying the attack using a POST request using the value *c:/* in the parameter *remember* on the URL *https://intranet.c4g-pt.eu/login*, the ZAP report an access to the folder *etc* but when we tried to replicate it, we were redirected to the login page. Also is important to highlight that this attack had 39 instances, meaning that ZAP was able to exploit this vulnerability in 39 different situations. However, it can classified as a false positive.

The vulnerability reported as medium refers to the X-Frame-Options header not being set and was also detailed in the previous section

Finally the wapiti tool reported again no vulnerabilities found.

To see if there is any false positives the definitions of ZAP where changed, setting the attack strength and threshold both to High. In this second round of tests, the GDPR questions were not answered and the results from NMAP were the same as the previous round of tests.

Moving then to the ZAP tests, the difference from the previous tests was the increase by one for the high and medium vulnerabilities. Regarding the high one, it was reported the vulnerability *Remote OS Command Injection*. Again when trying to replicate the instances reported using cURL, it did not returned the values as on the report. However, was allowed the send of data containing the commands *sleep* for example, which means that some data sanitizing practices are not being done.

Regarding the medium vulnerability changes, now besides the one reported on the previous test also reported the Buffer Overflow vulnerability. This is characterized for the possibility to write in adjacent memory space, when there is more data in a buffer than it can handle. In this case is detailed on the report that is a *Potential Buffer Overflow* because when trying to do the HTTP attack requests it returned error 500, meaning Internal Server Error

Also the wapiti was changed setting attack level was set to 2, that increases the attacks payloads even if there is no parameter present in each endpoint found. Also was set the option *-scan-force* to aggressive, that increases the requests sent, based on the input parameters an URL or form may have. Again it reported no results.

5.4 Results discussion

Beginning with the GDPR checklists, both the scenarios tested answered the same questions and since they collect personal data and use authentication mechanisms, the EPOS GNSS compliance level is better than the one from the C4G. Is also possible that some questions were not fully understood, which led to a negative answer. For example the question asking for usage of encryption, which certainly used. So, maybe instead of only doing one survey in each scenario, it should have been done another one to avoid this kind of errors. But in the end, comparing both results is possible to observe that a data subject has his data rights more guaranteed in

the EPOS GNSS environment than in the C4G, that can possibly have severe security issues.
So with the use of the suggestions, some changes can be made for them to achieve more compliance specially on the accountability principle, where is expected to demonstrate and show what was done in order to be in compliance. The security assessment tools, specifically from ZAP, reported some false positives vulnerabilities, meaning that besides that, there is the need to test manually for those vulnerabilities, especially the ones reported as high as medium. The remaining vulnerabilities were not critical, being most of them related to missing parameters on requests headers, which were mostly identified as a medium vulnerability, but can be easily solved. Also is not possible to say with 100% certain that increasing the parameters on the security tools, that it will detect more vulnerabilities or decrease the number of false positives. Regarding the NMAP output in both scenarios, which by coincidence are hosted in the same machine, pointed the vulnerabilities found on the services running. When checked on the link suggested they are actual and met the versions being used. With those links and CVE codes is possible to know some attack surfaces, which can be exploited.

Chapter 6

Conclusion and future work

This dissertation has considered the impact of GDPR, especially on companies with limited resources which affects their ability to implement and understand the changes needed to be compliant. We have seen the problems raised due to the ever increasing use of technology and personal data to develop solutions tailor made to each user. The appearance of GDPR providing protection to the data subjects that belong to the EU has obviously complicated business practices.

Throughout the study several problems regarding the GDPR were identified. The difficulty of extracting from the regulations the exact meaning of certain points was one of them. From the articles reviewed the idea of organizing the GDPR into seven principles was identified Next the most important points associated with each principle were identified in order to be able to develop a tool to help and guide business.

Also and since it is one the the principles, namely integrity and confidentiality, a set of tools responsible for finding security flaws in web applications was integrated into our procedures. The final result, was a tool encompassing the GDPR regulations summarized into a checklist with suggestions and a set of (at least) two tools for the security assessment. Based on the tool outputs a report detailing those results was created.

The first conclusion to be made is the fact that all the original goals of this dissertation were addressed. From the regulation points extracted to the suggestions, including a cookie scanner and a concise security assessment, compose all the answers to the goals that were defined from the beginning. Even though some of the articles studied did not contributed directly in the development of PADRES, they were extremely helpful in building a solid and strong base and from there build the application. Also gave a lot of knowledge about areas and new technologies, which will be certainly used in the future, so it can be embraced and understood successfully. This statement can also be applied to anyone who reads this dissertation. Even if it already as some knowledge in the area.

Regarding the extraction of legal requirements, the points for each principle and the corresponding suggestions and questions were extracted by someone with no expertise or training in the law, all the necessary documentation was available, accessible and understandable. None the less in order to provide this tool to any company, this part should be reviewed by someone with knowledge due to its complex constraints and hidden clauses. Some of them also require a posterior research to complement the information already obtained.

One of the downsides regarding the security assessments tools, was the time that they took to finish as well as the quality of the results due to some false positives. Some of the tests lasted for four hours. The other downside on the security tools, was the fact that Wapiti did not report anything even when the parameters were changed to make attacks more powerful. But since the results from ZAP were false positives, the report from Wapiti can be correct. The point of using two tools that basically do the same, was exactly to have redundancy. Also the NMAP report can give an overview of the infrastructure and the vulnerabilities associated with the services running on the ports found.

Concluding, the tool can be helpful when first approaching the the GDPR, but based on the

results, is not possible to rely entirely on it, requiring further information from specialized sources. The same also applies to the security assessments. Based on the report make the decisions necessary to investigate the vulnerabilities more exhaustively, if necessary. Also the tool should have been tested on other applications in order to have more results and from there conclude which are the principles where is observed a bigger compliance absence and also see which are the most common vulnerabilities found and match them on the OWASP top 10.
Based on that, for the future is important the have that data so more suggestions and points can be extracted in order to have a more expressive report. The ideal would be a system capable of extracting that information automatically even though was found on the studied articles that such tools already exist but do not give the expected result [6]. Another important feature, would be a functionality capable of calculate the anonymization level of a given data set based on the entropy level [14] or based on the algorithms introduced here [15]. Even though adding new security assessment tools is a difficult task because it needs to be done manually, the use of them such as Arachni can introduce more redundancy and possibly found more vulnerabilities, supported by an improved interface to allow the user to choose the definitions of such tools.

Bibliography

[1] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, p. 3–11, Jul. 1999. [Online]. Available: https://doi.org/10.1145/329124.329126 vii, 1

[2] 147 million social security numbers for sale: Developing data protection legislation after mass cybersecurity breaches. Accessed: June 25, 2020. [Online]. Available: https://ilr.law.uiowa.edu/print/volume-103-issue-6/147-million-social-security-numbers-for-sale-developing-data-protection-legislation-after-mass-cybersecurity-breaches/ vii, 1, 12

[3] Eu data protection directive. Accessed: June 25, 2020. [Online]. Available: https://epic.org/privacy/intl/eu_data_protection_directive.html vii, 1

[4] Rights of the data subject. Accessed: June 25, 2020. [Online]. Available: https://gdpr-info.eu/chapter-3/ viii, 10

[5] Chapter 2 - principles. Accessed: June 25, 2020. [Online]. Available: https://gdpr-info.eu/chapter-2/ viii, 11

[6] V. Ayala-Rivera and L. Pasquale, "The grace period has ended: An approach to operationalize gdpr requirements," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 08 2018, pp. 136–146. viii, xii, xix, 4, 11, 14, 16, 17, 21, 41, 64

[7] P. N. Otto and A. I. Anton, "Addressing legal requirements in requirements engineering," in *15th IEEE International Requirements Engineering Conference (RE 2007)*, Oct 2007, pp. 5–14. viii, 4

[8] Owasp foundation. Accessed: June 25, 2020. [Online]. Available: https://www.owasp.org/index.php/Main_Page viii, 2

[9] Owasp top 10 - 2017 - the ten most critical web application security risks. Accessed: June 25, 2019. [Online]. Available: https://owasp.org/www-project-top-ten/OWASP_Top_Ten_2017/ viii, 2

[10] What is owasp? what are the owasp top 10? Accessed: June 25, 2020. [Online]. Available: https://www.cloudflare.com/learning/security/threats/owasp-top-10/ viii, 2, 33

[11] Epos-ip, glass is a unique open access platform for earth sciences research. Accessed: June 25, 2020. [Online]. Available: https://www.epos-ip.org/about/what-epos ix, 3

[12] Epos glass-api-dashboard. Accessed: Jun 2, 2020. [Online]. Available: https://glass.epos.ubi.pt:8080/GlassFramework/ x, xi

[13] Openssh release notes. Accessed: May 24, 2020. [Online]. Available: https://www.openssh.com/releasenotes.html x, 55

[14] L. Oliveira, F. Pereira, R. Misoczki, D. Aranha, F. Borges, M. Nogueira, M. Wangham, M. Wu, and J. Liu, "The computer for the 21st century: present security privacy challenges," *Journal of Internet Services and Applications*, vol. 9, 12 2018. xii, 27, 34, 39, 64

[15] R. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” vol. 2010, 05 2005, pp. 217– 228. xii, 31, 64

[16] Gnss data and products. [Accessed: June 25, 2020]. [Online]. Available: https://www.epos-ip.org/tcs/gnss-data-and-products/news/glass-unique-open-access-platform-earth-sciences-research xvii, 3

[17] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, “Privacy issues and data protection in big data: A case study analysis under gdpr,” 11 2018. xvii, 2, 10, 31, 37, 39

[18] G. Nelson, “Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification,” 04 2015. xvii, 27

[19] B. Fung, k. Wang, R. Chen, and P. Yu, “Privacy-preserving data publishing: A survey of recent developments,” *ACM Comput. Surv.*, vol. 42, 06 2010. xvii, 28, 29, 31

[20] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, “A systematic comparison and evaluation of k-anonymization algorithms for practitioners,” *Transactions on Data Privacy*, vol. 7, pp. 337–370, 12 2014. xvii, 28, 31

[21] H. Hamidovic, J. Kabil, and E. Šehić, “Eu general data protection regulation (gdpr) - anonymisation and pseudonymisation in function of data protection,” 04 2019. xvii, 30

[22] “Introduction to angular concepts,” accessed: March 23, 2020. [Online]. Available: https://angular.io/guide/architecture xvii, 42

[23] N. Vollmer. (2018, Sep) Article 83 eu general data protection regulation (eu-gdpr). [Online]. Available: http://www.privacy-regulation.eu/en/83.htm 1

[24] . A comprehensive approach on personal data protection in the european union. Accessed: June 25, 2020. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0609:FIN:EN:PDF 2

[25] Recital 26, eu gdpr. Accessed: June 25, 2020. [Online]. Available: http://www.privacy-regulation.eu/en/r26.htm 2

[26] Metadata management and distribution system for multiple gnss networks. Accessed: Aug 15, 2020. [Online]. Available: https://gnss-metadata.eu/site/index 3

[27] Owasp top 10 application security risks - 2017. Accessed: June 25, 2020. [Online]. Available: https://owasp.org/www-project-top-ten/OWASP_Top_Ten_2017/Top_10-2017_Top_10 4, 32

[28] S. D. Warren and L. D. Brandeis, “Ethical issues in the use of computers,” D. G. Johnson and J. W. Snapper, Eds. Belmont, CA, USA: Wadsworth Publ. Co., 1985, pp. 172–183. [Online]. Available: http://dl.acm.org/citation.cfm?id=2569.2679 7

[29] Universal declaration of human rights. Accessed: June 25, 2020. [Online]. Available: https://www.un.org/en/universal-declaration-human-rights/ 7

[30] ’no one shall be subjected to arbitrary or unlawful interference with his privacy, home or correspondence, nor to unlawful attacks on his honor and reputation.’. Accessed: June 25, 2020. [Online]. Available: https://www.lawyersnjurists.com/article/‘no-one-shall-be-subjected-to-arbitrary-or-unlawful-interference-with-his-privacy-home-or-correspondence-nor-to-unlawful-attacks-on-his-honor-and-reputation-’/#_ftn14 7

[31] OECD, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 2002. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/9789264196391-en 7

[32] "Regulations, directives and other acts," June 2020. [Online]. Available: https://europa.eu/european-union/eu-law/legal-acts_en 7

[33] Lex access to european union law. Accessed: June 25, 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32002L0058 8

[34] What are pecr? Accessed: June 25, 2020. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-pecr/what-are-pecr/ 8

[35] Lex - 32002l0058 - en. Accessed: June 25, 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32002L0058&from=EN 8

[36] L. J. Trautman, "Corporate directorss and officerss cybersecurity standard of care: The yahoo data breach," *SSRN Electronic Journal*, 2016. 8

[37] "Intention to fine marriott international, inc more than £99 million under gdpr for data breach," accessed: June 25, 2020. [Online]. Available: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/07/intention-to-fine-marriott-international-inc-more-than-99-million-under-gdpr-for-data-breach/ 8

[38] B. Givens, "Identity theft: How it happens, its impact on victims, and legislative solutions: Testimony for u.s. senate judiciary subcommittee on technology, terrorism, and government information," Jul 2000. [Online]. Available: https://privacyrights.org/resources/identity-theft-how-it-happens-its-impact-victims-and-legislative-solutions-testimony-us 8

[39] "The main differences between the dpd and gdpr." [Online]. Available: https://seeunity.com/whitepapers/main-differences-dpd-gdpr/ 8

[40] "Regulation (eu) 2016/679 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)," Apr 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2254-1-1 9, 10, 13, 14

[41] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "Eu general data protection regulation: Changes and implications for personal data collecting companies," *Computer Law Security Review*, vol. 34, no. 1, pp. 134 – 153, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364917301966 10, 15, 16, 21

[42] N. SIMONCINI, "Data protection officer (dpo)," 2017, accessed: January 26, 2020. [Online]. Available: https://edps.europa.eu/data-protection/data-protection/reference-library/data-protection-officer-dpo_en 10

[43] Vital interests. Accessed: June 25, 2020. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/vital-interests/ 11

[44] (2018) Guidelines on consent under regulation 2016/679. Accessed: June 25, 2020. [Online]. Available: http://odo.uw.edu.pl/wp-content/uploads/sites/278/2018/01/Wytyczne-w-sprawie-zgody-na-mocy-rozporza%CC%A8dzenia-2016679-WP-259-EN.pdf 11

[45] The personal data processing principle of data minimization. Accessed: June 25, 2020. [Online]. Available: https://www.i-scoop.eu/gdpr/gdpr-personal-data-processing-principles/#The_personal_data_processing_principle_of_data_minimization 11

[46] How can you practise data minimisation? Accessed: June 25, 2020. [Online]. Available: https://www.experian.co.uk/business/glossary/data-minimisation/ 12

[47] Definitions. Accessed: June 25, 2020. [Online]. Available: http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm 12

[48] "Third countries," accessed: June 25, 2020. [Online]. Available: https://gdpr-info.eu/issues/third-countries/ 12

[49] K. Julisch and K. Julisch, "Gdpr-accountability principle," 2018, accessed: January 26, 2020. [Online]. Available: https://www2.deloitte.com/ch/en/pages/risk/articles/gdpr-accountability-principle.html 13

[50] "Individual rights," accessed: June 25, 2020. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/ 13

[51] "Directive 95/46/ec 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," Oct 1995. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML 13

[52] "29148-2011 - iso/iec/ieee international standard - systems and software engineering – life cycle processes –requirements engineering." [Online]. Available: https://standards.ieee.org/standard/29148-2011.html 15, 16

[53] A. Davis, S. Overmyer, K. Jordan, J. Caruso, F. Dandashi, A. Dinh, G. Kincaid, G. Ledeboer, P. Reynolds, P. Sitaram, A. Ta, and M. Theofanos, "Identifying and measuring quality in a software requirements specification," in *[1993] Proceedings First International Software Metrics Symposium*, May 1993, pp. 141–152. 16

[54] T. Breaux, *Introduction to IT privacy: a handbook for technologists*. International Association of Privacy Professionals, 2014. 16

[55] Y. Martin and A. Kung, "Methods and tools for gdpr compliance through privacy and data protection engineering," in *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, April 2018, pp. 108–111. 16, 31

[56] S. Wachter, "Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr," *Computer Law Security Review*, vol. 34, no. 3, p. 436–449, 2018. 17, 21

[57] M. Barati and O. Rana, "Enhancing user privacy in iot: Integration of gdpr and blockchain," in *Blockchain and Trustworthy Systems*, Z. Zheng, H.-N. Dai, M. Tang, and X. Chen, Eds. Singapore: Springer Singapore, 2020, pp. 322–335. 19, 21

[58] X. Larrucea, M. Moffie, S. Asaf, and I. Santamaria, “Towards a gdpr compliant way to secure european cross border healthcare industry 4.0,” *Computer Standards Interfaces*, vol. 69, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920548919304544 20, 21

[59] K. Schweichhart, “Reference architectural model industrie 4.0 (rami 4.0),” *An Introduction. Available online: https://www. plattform-i40. de I*, vol. 40, 2016. 21

[60] “Preparing your organisation for the general data protection regulation - your readiness checklist,” accessed: March 4, 2020. [Online]. Available: https://www.dataprotection.ie/sites/default/files/uploads/2019-04/A-Guide-to-help-SMEs-Prepare-for-the-GDPR.pdf 22

[61] “Gdpr checklist for data controllers,” accessed: March 4, 2020. [Online]. Available: https://gdpr.eu/checklist/ 22

[62] “Controllers checklist,” accessed: March 4, 2020. [Online]. Available: https://ico.org.uk/for-organisations/data-protection-self-assessment/controllers-checklist/ 22

[63] Eu gdpr readiness assessment tool. Accessed: Jun 25, 2020. [Online]. Available: https://advisera.com/eugdpracademy/eu-gdpr-readiness-assessment-tool/ 22

[64] “How microsoft tools and partners support gdpr compliance,” accessed: March 4, 2020. [Online]. Available: https://www.microsoft.com/security/blog/2017/12/19/how-microsoft-tools-and-partners-support-gdpr-compliance/ 22

[65] Snow gdpr risk assessment. Accessed: Jun 25, 2020. [Online]. Available: https://www.snowsoftware.com/int/products/snow-gdpr-risk-assessment 22

[66] C. Christmann, J. Falkner, A. Horch, and H. Kett, “Identification of it security and legal requirements regarding cloud services,” in *IEEE CLOUD 2015*, 2015. 22

[67] G. Boella, L. Humphreys, R. Muthuri, P. Rossi, and L. van der Torre, “A critical analysis of legal requirements engineering from the perspective of legal practice,” in *2014 IEEE 7th International Workshop on Requirements Engineering and Law (RELAW)*, 2014, pp. 14–21. 22

[68] H. Gjermundrod, I. Dionysiou, and K. Costa, “privacytracker: A privacy-by-design gdpr-compliant framework with verifiable data traceability controls,” vol. 9881, 06 2016, pp. 3–15. 23

[69] A. Tsohou, E. Magkos, H. Mouratidis, G. Chrysoloras, L. Piras, M. Pavlidis, J. Debussche, M. Rotoloni, and B. G.-N. Crespo, “Privacy, security, legal and technology acceptance elicited and consolidated requirements for a gdpr compliance platform,” *Information & Computer Security*, 2020. 23

[70] O. Akhigbe, D. Amyot, and G. Richards, “A systematic literature mapping of goal and non-goal modelling methods for legal and regulatory compliance,” *Requirements Engineering*, vol. 24, no. 4, pp. 459–481, 2019. 23

[71] F. T. C. Editors. (2018, Dec) Data privacy vs. data protection: Understanding the distinction in defending your data. Accessed: June 25, 2020. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2018/12/19/data-privacy-vs-data-protection-understanding-the-distinction-in-defending-your-data/#7fa5cc2050c9 25

[72] C. Gates, “Access control requirements for web 2.0 security and privacy,” 01 2007. 25, 39

[73] D. Chen and H. Zhao, “Data security and privacy protection issues in cloud computing,” in *2012 International Conference on Computer Science and Electronics Engineering*, vol. 1, March 2012, pp. 647–651. 25

[74] V. R. Q. Leithardt, “Ubipri – middleware para controle e gerenciamento de privacidade em ambientes ubíquos,” Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, 2015. 26

[75] D. Zhang, “Big data security and privacy protection,” in *8th International Conference on Management and Computer Science (ICMCS 2018)*. Atlantis Press, 2018/10. [Online]. Available: https://doi.org/10.2991/icmcs-18.2018.56 26

[76] R. L. Rivest and R. D. Silverman, “Arestrong’primes needed for rsa?” in *IN THE 1997 RSA LABORATORIES SEMINAR SERIES, SEMINARS PROCEEDINGS*, 1999. 26

[77] A. Shah, V. Banakar, S. Shastri, M. Wasserman, and V. Chidambaram, “Analyzing the impact of GDPR on storage systems,” *CoRR*, vol. abs/1903.04880, 2019. [Online]. Available: http://arxiv.org/abs/1903.04880 26

[78] C. Gentry, “Computing arbitrary functions of encrypted data,” *Commun. ACM*, vol. 53, no. 3, p. 97–105, Mar. 2010. [Online]. Available: https://doi.org/10.1145/1666420.1666444 26

[79] J. Holvast, “History of privacy,” in *The Future of Identity in the Information Society*, V. Matyáš, S. Fischer-Hübner, D. Cvrček, and P. Švenda, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 13–42. 27

[80] T. Menzies, E. Kocaguneli, B. Turhan, L. Minku, and F. Peters, *Sharing data and models in software engineering*. Morgan Kaufmann, 2014. 27, 28, 29

[81] F. Kerschbaum, “Privacy-preserving computation,” in *Privacy Technologies and Policy*, B. Preneel and D. Ikonomou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 41–54. 27

[82] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, “Privacy-preserving data publishing,” *Foundations and Trends in Databases*, vol. 2, pp. 1–167, 01 2009. 28

[83] Y. S.ALDEEN and M. Salleh, “Privacy preserving data utility mining architecture,” pp. 253–268, 01 2019. [Online]. Available: https://doi.org/10.1016/B978-0-12-815032-0.00018-4 29

[84] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving data mining,” *SIGMOD Rec.*, vol. 33, no. 1, p. 50–57, Mar. 2004. [Online]. Available: https://doi.org/10.1145/974121.974131 29

[85] A. Sharma, P. Panday, R. Baghel, and P. Saini, “A survey on techniques for privacy preserving data publishing (ppdp),” *MIT International Journal of Computer Science and Information Technology*, vol. 4, no. 2, pp. 60–64, 2014. 29

[86] M. Hintze, “Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency,” *International Data Privacy Law*, vol. 8, no. 1, pp. 86–101, 12 2017. [Online]. Available: https://doi.org/10.1093/idpl/ipx020 29

[87] S. Stalla-Bourdillon and A. Knight, “Anonymous data v. personal data-false debate: An eu perspective on anonymization, pseudonymization and personal data,” *Wis. Int’l LJ*, vol. 34, p. 284, 2016. 29, 30

[88] Lawfulness of processing. Accessed: February 27, 2020. [Online]. Available: https://gdpr-info.eu/art-6-gdpr/ 30

[89] Definitions. Accessed: February 27, 2020. [Online]. Available: https://gdpr-info.eu/art-4-gdpr/ 30

[90] “Pseudonymisation techniques and best practices,” Dec 2019, accessed: February 28, 2020. [Online]. Available: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices 30

[91] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019. 30

[92] “Opinion 05/2014 on anonymisation techniques,” accessed: March 3, 2020. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/index_en.htm 30, 31

[93] L. Bolognini and C. Bistolfi, “Pseudonymization and impacts of big (personal/anonymous) data processing in the transition from the directive 95/46/ec to the new eu general data protection regulation,” *Computer Law Security Review*, vol. 33, no. 2, pp. 171 – 181, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364916302151 31

[94] L. Tarhonen. Pseudonymisation of personal data according to the general data protection regulation. Accessed: February 27, 2020. [Online]. Available: https://www.edilex.fi/artikkelit/18073.pdf 31

[95] “Privacy by design gdpr,” accessed: March 4, 2020. [Online]. Available: https://www.privacytrust.com/gdpr/privacy-by-design-gdpr.html 31

[96] “Gdpr privacy by design made simple,” accessed: March 4, 2020. [Online]. Available: https://www.privacytrust.com/gdpr/gdpr-privacy-by-design-made-simple.html 31

[97] S. Danon, “Gdpr-privacy by design and by default,” accessed: March 4, 2020. [Online]. Available: https://www2.deloitte.com/ch/en/pages/risk/articles/gdpr-privacy-by-design-and-by-default.html 31

[98] A. Cavoukian *et al.*, “Privacy by design: The 7 foundational principles,” *Information and privacy commissioner of Ontario, Canada*, vol. 5, 2009. 32

[99] G. Danezis, J. Domingo-Ferrer, M. Hansen, J. Hoepman, D. L. Métayer, R. Tirtea, and S. Schiffner, “Privacy and data protection by design - from policy to engineering,” *CoRR*, vol. abs/1501.03726, 2015. [Online]. Available: http://arxiv.org/abs/1501.03726 32

[100] D. Saraiva, V. Leithardt, D. de Paula, M. A. Sales, G. González, and P. Crocker, “Prisec: Comparison of symmetric key algorithms for iot devices,” vol. 19(19), p. 4312, 2019. 35, 39

[101] W. Itani, A. Kayssi, and A. Chehab, “Privacy as a service: Privacy-aware data storage and processing in cloud computing architectures,” in *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, Dec 2009, pp. 711–716. 36, 39

[102] “The principles,” accessed: March 12, 2020. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/ 41

[103] R. T. Fielding, “Architectural styles and the design of network-based software architectures. 2000,” *University of California, Irvine*, p. 162, 2000. 42

[104] “Http request methods,” accessed: March 20, 2020. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods 42

[105] “Http response status codes,” accessed: March 20, 2020. [Online]. Available: https://developer.mozilla.org/pt-PT/docs/Web/HTTP/Status 42

[106] C. KarbaliotisFollowCounsel and C. KarbaliotisCounsel, “The nightmare letter: A subject access request under gdpr.” [Online]. Available: https://www.linkedin.com/pulse/nightmare-letter-subject-access-request-under-gdpr-karbaliotis/ 50

[107] “Chapter 1. getting started with nmap,” accessed: May 5, 2020. [Online]. Available: https://nmap.org/book/intro.html 50

[108] Accessed: May 5, 2020. [Online]. Available: https://pypi.org/project/python-nmap/ 50

[109] Accessed: May 24, 2020. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/X-Frame-Options 56

Appendix A

A.1 Images supporting the scenario description

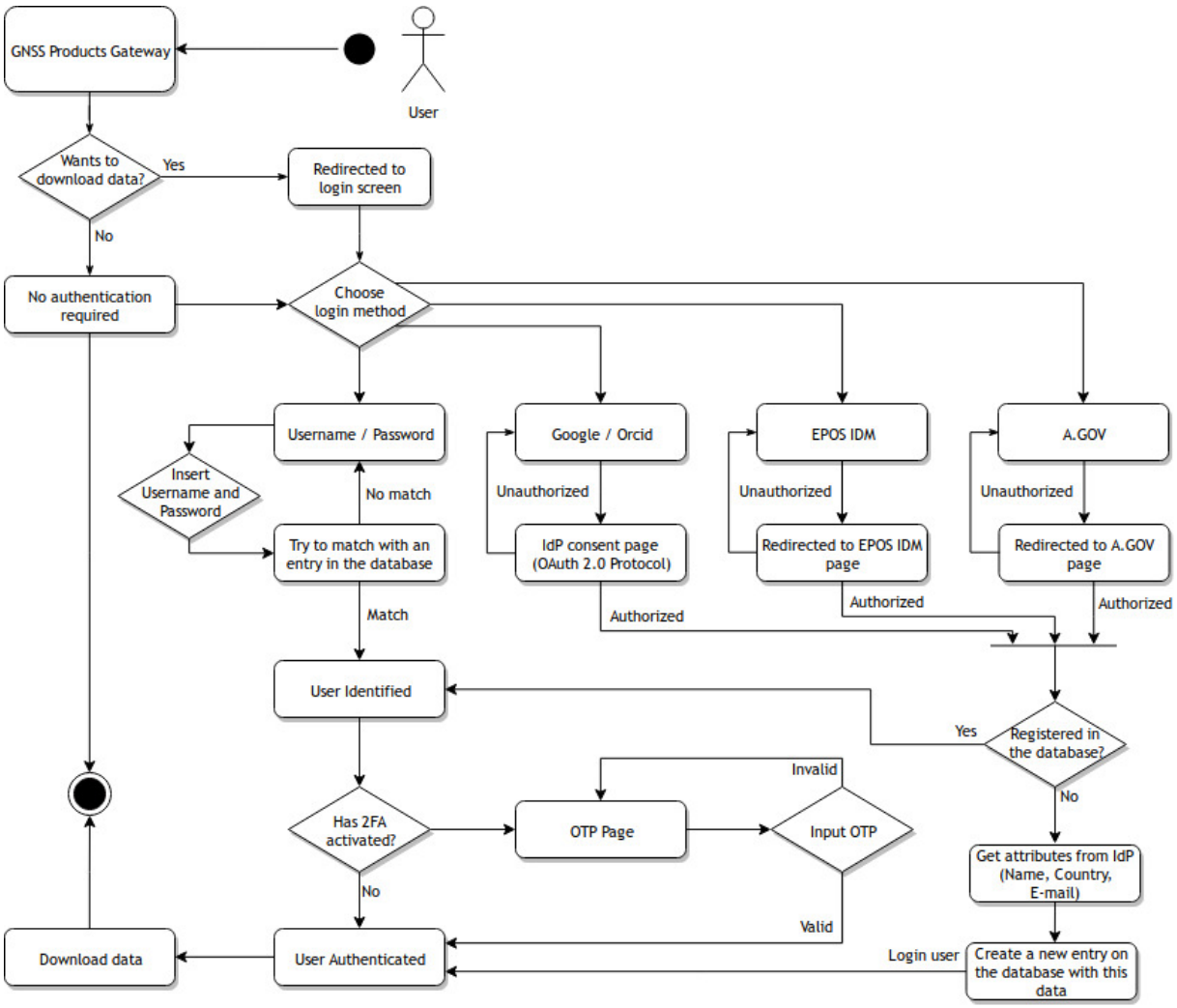


Figure A.1: Authentication workflow

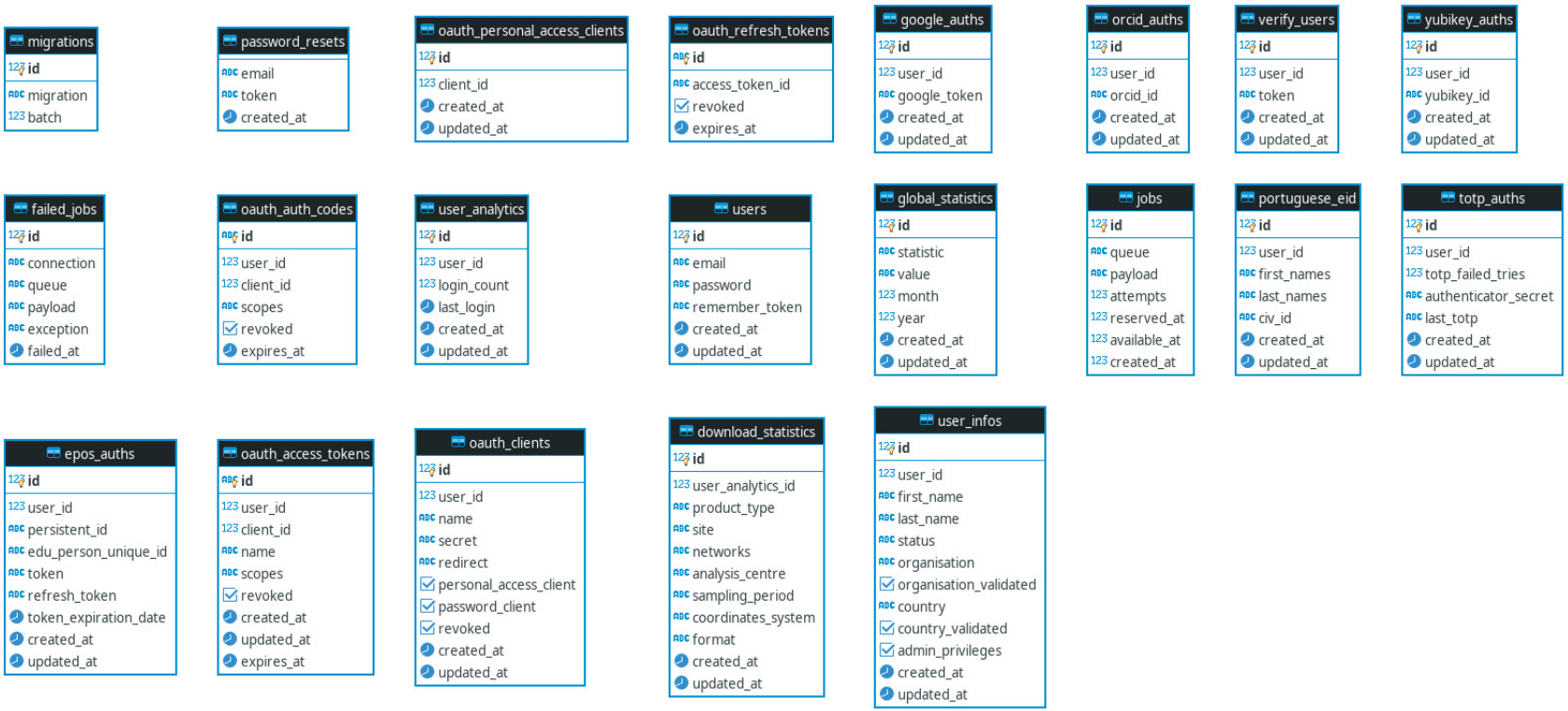


Figure A.2: Authentication database schema

A.2 GDPR questions and suggestions

GDPR report for C4G Intranet following the Portugal's specific rules

Lawfulness, fairness and transparency

In compliance with:

- Does your software ask and record for consent?
- Does the consent inform the Inviduals about the processing objectives?
- Does your application provide any information regarding the Individual's rights?

Not in compliance with:

- Have you performed any audit to map data flows?

  Suggestions to be in compliance

  - You need the follow the data flow in order to understand what and where data is collected.
  - Where data is stored is also important to analyze possible security problems.

- Does your software record a users consent choices ?

  Suggestions to be in compliance

  - You need to be able to demonstrate the data subject's consent

Purpose limitation

In compliance with:

- Is Personal data being collected for the specified, explicit and legitimate purpose?
- Do the Individuals have access to the purpose details?
- Do you regularly update your purpose based on the changes made for the processing?

Not in compliance with:

- Do you ask for new consent if the purpose changes?

  Suggestions to be in compliance

  - Data from one purpose can be used in other situations rather than one specified in the consent only if the new purpose is compatible with the original purpose

Data minimisation

In compliance with:

- The data being collected is sufficient to fulfill the consent purposes
- Is your application only holding the data being used?
- Do you regularly review the data stored?
- Can you still achieve your purpose, if the data collected is reduced?
- Do you clear data, when it is no longer needed?

Not in compliance with:

- Is it possible to demonstrate your data minimization practices?

  Suggestions to be in compliance

  - You can determine the absolutely necessary data to fulfill your purpose
  - Include standardized policies to clean the data, for example deleting outdated data can not only improve you

Figure A.3: GDPR questions 1

data minimization adoption but also reduce security risks

Accuracy

In compliance with:

- Does your application provide mechanisms to keep data updated?
- Do you inform individuals about their right of rectification?

Not in compliance with:

- Are you aware of your right to refuse requests for rectifications?

  Suggestions to be in compliance

  - You can decline their request if it is manifestly unfounded and excessive requests.

- Do you comply with the limit of one month to answer requests to update data(right of access)?

  No suggestions available

- Does your application erase incorrect data?

  No suggestions available

Storage limitation

In compliance with:

- Are you aware that you can keep data for longer than needed if you are only keeping it for public interest archiving, scientific or historical research, or statistical purposes.

Not in compliance with:

- Is it possible to justify the time frame for the retained data?

  No suggestions available

- Does your application automatically deletes the data after the time frame expires?

  No suggestions available

- Does your application provides a way, so the individual can erease his data(right to erasure)?

  No suggestions available

- Do you have any mechanisms to anonymize data?

  Suggestions to be in compliance

  - Anonymizing data allows you to keep the data, after the time frame defined ends.

Integrity and confidentiality (security)

In compliance with:

- Are you aware that the security of personal data is the data controller's responsibility?
- Do you have any measures regarding any data leak?

Not in compliance with:

- Do you apply techniques to protect against unlawful and unauthorised processing?

Figure A.4: GDPR questions 2

Suggestions to be in compliance

- Defining user permissitions to access data, can prevent those problems.

- Do you have any mechanisms to pseudonymize data?

Suggestions to be in compliance

- Is a data management technique, where the data controller swaps the individualâ€™s direct identifiers, such as email or phone number, with a pseudonym. Then the data processor can process the data without exposing the sensitive data. Then when data goes back to the data controller, he can rebuild the original data through re-identification techniques

- Does your application use encryption?

No suggestions available

Accountability

In compliance with:

Not in compliance with:

- Can you demonstrate compliance with the points answered before?

No suggestions available

Rules Specific for the selected country

No principles defined

The software C4G Intranetdoes not comply with 15 rules from a total of 31

Figure A.5: GDPR questions 3