TÉCNICO LISBOA

BPML Tool for GDPR Compliance Checking

Rodrigo de Bragança Santa Clara Reis

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Carlos Nuno da Cruz Ribeiro
Eng. Filipe Apolinário

Examination Committee

Chairperson: Prof. Teresa Maria Sá Ferreira Vazão Vasques
Supervisor: Prof. Carlos Nuno da Cruz Ribeiro
Member of the Committee: Prof. Hugo Miranda

June 2019

“I will be remembered by less people than a gorilla when I die”

u/mourato

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

The research for this paper was financed by the European Commission under Horizon 2020 grant agreement no. 740712 (COMPACT-COmpetitive Methods to protect local Public Administration from Cyber security Threats). This paper reflects the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Resumo

Nos últimos anos temos assistido ao processamento e armazenamento em massa dos nossos dados pessoais, devido à grande disponibilidade de serviços digitais, pondo em risco a nossa privacidade. Como consequência, vários países iniciaram uma reforma das leis e regulamentos relacionados com a proteção de dados, para que os seus cidadãos possam usufruir dos serviços online mantendo a sua privacidade. Liderando a reforma, o Regulamento Europeu de Proteção de Dados (RGPD) impõe regras mais rígidas sobre o tratamento dos dados pessoais de cidadãos europeus.

As organizações precisam de uma ferramenta que lhes permita garantir que este tratamento está em conformidade com o regulamento, no entanto, as ferramentas de monitorização padrão não estão orientadas para o tratamento de dados pessoais. Esta dissertação propõe o uso de uma ferramenta de deteção de intrusão baseada em processos de negócio, para monitorizar o tratamento de dados, de maneira a que seja possível a verificação da conformidade com o RGPD. O uso de uma linguagem gráfica de modelação de processos de negócios, juntamente com a ferramenta BP-IDS, um sistema de deteção de intrusão que nos permite conciliar conceitos de proteção de dados e os processo de negócio, ajuda a organização a especificar processos que lidam com dados pessoais.

A proposta apresentada é avaliada no contexto do projecto COMPACT, em parceria com o INOV e o Município da Amadora. É implementado um caso de uso para o direito ao esquecimento e, adicionalmente, a conformidade com os princípios de limitação do armazenamento e limitação da finalidade é monitorizada.

Palavras-chave: RGPD, proteção de dados, conformidade com RGPD, monitorização de dados pessoais, monitorização de processos de negócios, segurança de dados pessoais

Abstract

In recent years we have been watching our privacy becoming at risk in mass data collection, analysis, processing and storing: all made possible by social networks and due to a large availability of digital services. As a result, countries are reforming their legal acts on data protection, so that individuals may keep their privacy while maintaining their access to the usual online services. Leading the reform, the new European General Data Protection Regulation (GDPR) calls for some stiffer rules on data protection for all personal data on European citizens.

Organizations need a tool that allows them to assure their processing is compliant and current monitoring applications are not data privacy driven besides lacking automated auditing methods. This dissertation proposes the use of a business process intrusion detection tool to monitor data protection concerns over an organization processes, in a manner that allows for compliance verification. The use of business process modeling languages helps organizations to specify their processes that handle personal data and BP-IDS, a Business Process Intrusion Detection System, introduces data protection concepts over those baseline processes. This approach validation is integrated within the context of the COMPACT project, partnering INOV and the Municipality of Amadora (CMA). A right to be forgotten scenario use case is implemented and the conformity to the principles of *storage minimization* and *purpose limitation* is monitored.

Keywords: GDPR, Data Protection, GDPR compliance, Data monitoring, Business Process monitoring, Purpose, Cyber-security

Contents

List of Tables

List of Figures

Nomenclature

BP-IDS Business Process Intrusion Detection System

BPMN Business Process Model and Notation

DPO Data Protection Officer

GDPR General Data Protection Regulation

IDS Intrusion Detection System

Chapter 1

Introduction

The word *privacy* in the cybersecurity context is recurrently appearing in the media and increasingly getting people's attention, result of several data leaks in past years, being *Cambridge Analytica* scandal [1] among the most notorious examples. Furthermore, the enactment of the European General Data Protection Regulation introduces new rules for those who process personal data, in an attempt to improve European citizen's privacy.

Despite late cybersecurity incidents, concerns and legislation about privacy and data protection started years back [2]. However, former legislation and efforts, take for example, the 1995 Data Protection EU Directive [1] , were not unified at all as several countries implemented specific laws when needed, with the consequence that each country in EU had different levels of laws regarding data and not always compatible among all member states [3]. Additionally, the digital world is in constant evolution and in recent years we witnessed to significant changes on how data is collected and handled, aggravated with the use of machine learning and big data analytics, which resulted in mass data collection, analysis, processing and storing made possible by social networks and large availability of digital services. Uncontrolled, these systems represent a risk for privacy(as data became available in distributed systems, and consequently, susceptible to correlation) and other legal issues provoked by automated individual decision-making, including profiling. As a result of such collecting of personal data, countries have been reforming their legal acts on data protection taking into consideration this technological progress. The data privacy laws have to be addressed so that individual persons may keep their privacy while maintaining their access to usual online services. Leading the reform, due to its extensive impact on the industry, is the European General Data Protection Regulation.

Before continuing, and within the scope of this work, the following concepts and terms should be considered. *Privacy* is a human right by the Council of Europe in the Convention for the Protection of Human Rights and Fundamental Freedoms [4] and is defined as the right to respect for every person's private and family life, home and correspondence. This term is widely used and not to be confused with *Data Privacy* (*Information Privacy*) or *Data Protection*. These two are also distinct from each other: whereas *Data Privacy* relates to the conception of laws governing the collecting and processing of

[1] 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, OJ L 281 of 23.11.95.

personal data, *Data Protection* refers to the implementation of measures to protect data subject's rights when collecting and processing the data [5–7].

1.1 Motivation

As previously mentioned data protection has always been an issue, hence some organizations have already implemented security measures to protect data and comply with legal requirements. Even if that is not the case, some industry standards provide a good starting point, as for instance, the case of *ISO 27001* compliant organizations [8]. However, the main challenge with GDPR is that they have to be fully compliant, and have to be able to show they are, otherwise some heavy fines are up for those who do not abide. For those organizations that have their full system built within the same proprietary ecosystem, most software product companies adapt their applications to be able to demonstrate compliance. Some of them with advanced monitoring tools that allow users to visualize and control the data, activities and logs that are managed by the system [9]. On the other hand, for the ones that do not have such ecosystems, changing their long term installed software can be a challenge. This can be particularly complex for small and medium enterprises (SMEs), who may not have the personnel nor expertise to endeavor such a project.

Showing "compliance" is essentially to assure that the regulation requirements are fulfilled. Entities that process personal data have to show their compliance with the data protection regulation. Although they may keep track of all actions over their personal data, they need a tool that can verify if there is any behavior that deviates from the expected data flow. The tools that can detect violations to the normal behavior of a system are Intrusion Detection Systems (IDSs). These tools monitor the activities of a system in order to detect illegal actions performed on it, and report the violations, in a structured and consistent way, to the security officer. IDSs detection methods can be based on known attacks signatures, baseline specifications of the system or even using machine learning methods to learn the system behavior beforehand (anomaly detection).

This dissertation proposes monitoring personal data using an IDS, and shows that a set of processes that handle personal data are compliant with GDPR. However, not all IDSs are considered suitable for this quest as signature detection can only detect attacks to which a signature has already been produced, while anomaly detection usually suffer from a high number of false alarms, and can model attack behavior as acceptable in the learning phase [10]. Specification based IDSs main disadvantage is the manual specification of the system expected behavior by the system administrator. Due to GDPR imposing strict rules and restrictions over personal data processing, organizations must have their system processes well defined and structured when processing data. That means a specification of those processes, the system expected behavior, should be within reach, which creates the opportunity for specification based IDSs to be used.

Business Process Intrusion Detection System (BP-IDS) [2] compares a predefined specification of the system business processes, the baseline, to the real-time state of the system, which is monitored by

[2] See BP-IDS homepage in [11]

passive sensors. The initial model should represent the expected behaviour of every process, and, any deviation represents a violation and the administrator is notified. However, due to the fact that business process related to personal data are being monitored, it becomes difficult to model straightforward GDPR compliant baselines. In order to delineate these, an easy to learn and use graphical language is required. BP-IDS is an intrusion detection solution that directly allows BPMN models of the "normal" system's processes and captures basic chunks of communications to construct the system state in real time and diagnose for deviations or non-compliances. It is crucial to establish that within the context of this work only processes that in any way process personal data are considered.

In this dissertation an attempt is made to apply this solution, monitoring personal data related processes and detecting any irregularity in the processing of personal data. The validation is integrated with the context of the COMPACT project [3], partnering INOV [4] and the Municipality of Amadora [5]. A right to be forgotten scenario use case is implemented and the principles of storage minimization and purpose limitation are monitored.

1.2 Topic Overview

1.2.1 General Data Protection Regulation

The EU General Data Protection Regulation (GDPR) [6] is the new European data protection legislation that entered into force on 25 May 2018. Adopted on 27 April 2016, it replaces the Data Protection Directive 95/46/EC and was it designed to empower all EU citizens regarding their data privacy and 'harmonize' laws and organizations' approaches on data privacy across Europe. The European regulation affects all member states directly and it is not required to be transposed into their own laws, which comes to solve the fact that the previous data protection directive had become fragmented across the EU [3]. The goal, besides assuring data subject's rights, is to foster the Digital Single Market [7], only possible through the guarantee of free flow of personal data between EU Member States and reinforced trust and security of consumers, both provided by this new regulation [15]. It's worth mentioning that European regulation applies when processing data of European citizens, which means it will not only affect EU member countries, but also all organizations that have services with European subjects.

1.2.2 Processing of personal data

The *personal data* concept is defined as *"any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in*

[3] See COMPACT Project homepage in [12]
[4] See INOV homepage in [13]
[5] See Municipality of Amadora homepage in [14]
[6] 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), OJ L 119, 4.5.2016.
[7] *A Digital Single Market is one in which the free movement of goods, persons, services and capital is ensured and where individuals and businesses can seamlessly access and exercise online activities under conditions of fair competition, and a high level of consumer and personal data protection, irrespective of their nationality or place of residence* [15].

particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" [16, art. 4]. Noteworthy to say that data can be in any format, both paper or electronically stored and, moreover, it should be considered as personal data (within the scope of the law) in all cases where the data subjects can be identified, even in those cases where data that has been de-identified, encrypted or pseudonymized. Furthermore, in this work context the term *data* is always referring to the concept of personal data.

According to the data regulation [16] *processing* means *" any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction"* . When referring entities that process personal data, GDPR makes the distinction between data controllers and data processors. Data controller is the system that collects and handles the data, while data processor processes data in behalf of the controller and respecting the privacy policies set by the former. This document will consider all the interactions with a data controller assuming it also processes data. The interactions between entities are explained on the chapter 4.1.2

The foundations of data protections regulations are based on principles. These principles, perfected by previous directives and regulations throughout data protection history, are defined by GDPR [16, Article 5]. A summary of each principle, and direct data subject rights are presented:

- **Lawfulness, Fairness and Transparency** Collecting and processing personal data must be justified by contractual or legal reasons, legitimate interests of the controller, public interest or must have explicit consent from the data subject, consent that must be given voluntarily and in a non-passive manner. Data subjects have the right to be informed of the use of their data, as well as, the right to access said data and the right to portability in an interoperable format transmissible to other controllers.

- **Purpose Limitation** The *purpose* for which data is collected has to be respected by whoever processes data. The *purpose* for which the consent is given must be explicit and easily understandable to the person consenting. The data subject must have the right to easily withdraw the consent previously given, and the right to object to the processing if the data is used for other *purposes*.

- **Data Minimization** The data that is collected must be limited to what is necessary and relevant to the *purpose* of processing.

- **Accuracy** The controller or the processor who keeps the data has to guarantee accuracy, thus requiring regular updates if still needed for the collected *purpose*. The data subject has the right to the rectification of his data and also to object in case of incorrectness.

- **Storage Limitation** Data should only be kept if it is still needed for the *purpose* of collection and erased when no longer needed. The data subject has the right to erasure, which means have

the possibility of erasing all his personal data. There are some exceptions and this right can be rejected due to contractual or legal reasons, or due to legitimate interests of the controller or by public interest.

- **Integrity and Confidentiality** The controller is responsible of applying appropriate technical or organizational measures in order to follow all the principles, while maintaining the data secure. Data protection concepts should be integrated as soon as possible into the development phase of technologies and these must be optimized by default for the maximum privacy possible.
- **Accountability** This last principle increases the burden to data controllers and processors, who besides complying with the regulation are now responsible for, and must be able to demonstrate their compliance to the principles. This responsibility is even greater where processing has a high risk of impact on the data subject rights. Furthermore, it can lead to the need of measures as risk/impact assessment and keeping a record of all the logs or even defining a Data Protection Officer. To enforce this GDPR gives supervisory authorities the power to impose fines on controllers and processors.

1.3 Thesis Contributions

Thesis question is a proposition to use intrusion detection technology to monitor and show compliance with GDPR and can be formulated as follow:

> *"Can process business process intrusion detection technology be used to monitoring data protection concerns over an organization's processes in a manner that allows GDPR compliance verification."?*

The main contributions here provided will attempt to answer these questions and provide an approach on how it can be accomplished using INOV's BP-IDS. The main points addressed will be:

- Identify main data protection requirements and concepts that should be monitored for verifying compliance.
- Select the personal data related processes, responsible for personal data flows, and model them as a BPMN specification.
- Assure BP-IDS can monitor personal data related processes and cross check the system current state against the business process specification, following all the rules, detecting any non-compliance.

1.4 Thesis Outline

The content of this dissertation is divided into research, implementation and validation, being distributed over the three next chapters. An insight on each chapter will be described in the remaining of this section.

The research for this work, reviewed in Chapter 2, progresses through several areas that offer solutions related to the enhancing of privacy over user data and the enforcing of data protections policies. These areas are grouped into data protection translation and ontologies, business process graphical modeling languages applied to data protection, privacy enhancing technologies and purpose-based access control. Chapter 3 briefly describes BP-IDS, the configurations needed and how it is expected to work. Chapter 4 introduces a solution that monitors compliance to data protection regulations using business process intrusion detection technology. The chapter starts by defining what are the data protection concepts and requirements to be monitored. To be able to verify that data protection requirements are being met, an explanation is provided on how to model a BPMN specification of the business processes related to personal data, followed by the description of the BP-IDS configurations needed to monitor the data protection concepts: data, purpose, rules and restrictions. Chapter 5 validates the proposed solution. The evaluation is based on an use case scenario which applies the right to erasure. The scenario of the COMPACT trials implemented in the CMA premises, permitted the evaluation of this solution in a real world system architecture and data. The several tests are divided into three parts. The first part evaluates if the BPMN specification of the system processes is reliable with the reality. The second part exploits the data protection concepts to demonstrate that it is possible to detect violations to data protection regulations. Finally the third part presents a tryout in performance when an infringement is detected and general applicability. Chapter 6 concludes the dissertation and presents the future work.

Chapter 2

Related Work

The world of laws, underlying in regulations, and the world of IT systems exist in very different perspectives, being fundamental to make a connection between them if one wishes to model compliant systems. Thus, it is important to find ways to transform written human language to models or to computer languages that can be later implemented or executed. When it comes to data protection and privacy, this comes down to implementing security requirements and data protection measures.

This chapter reports several proposals that influenced or contributed to the ideas generated in this document. The remainder of the chapter assumes the following structure: Section 2.1 presents proposals to the main constituents in the data protection context and how data protection requirements can be extracted from regulations into models and ontologies [1], so they can be implemented or enforced; Section 2.2 shows the usefulness of using business processes in data protection compliance monitoring; On Section 2.2.1, several studies introduce extensions to business processes graphical languages to express security concerns into business processes; Section 2.3 briefly shows how multiple security enhancing tools can be integrated with the purpose of complying with data protection regulations; Section 2.4 builds a purpose-based access control model, which is based on several works that build on each other; To sum up, Section 2.5 makes an evaluation on the knowledge gathered and how it contributes to this thesis.

2.1 From regulations to models

Before modeling security requirements is is important to define the main actors that affect the data protection domain. This way, it is easier to identify the responsibilities that each of them have when complying with the regulation requirements.

[19] created a GDPR-standards compatible software incorporation case, combining regulation and information privacy into UML use case diagrams. This paper selects the six actors based in the GDPR:

- Data Subject

[1] The organization of knowledge on a particular domain of study, which provides a framework that improves communication and inter-operability is referred to as an ontology [17, 18]

- Data Processor
- Data Controller
- Data Protection Officer (DPO, person or legal entity)
- Supervisory Authority (SA)
- Third party (person or legal entity)

Following the description of actors in data protection context in the remainder analyzed papers [20–22], [19] assumed the actors as defined in GDPR. Similarly, this work will focus in the actors as defined by the GDPR.

On the data protection context, focus is given to the duties of the data controllers and to the rights of the data subjects. Data subject rights are based on data protection principles [19, 20, 23] and the duties of the data controller correspond to the responsibility to apply and comply with the rules on data protection. There are several approaches it when comes to defining the data controller duties based on a data protection regulation.

[20] defines an ontology to provide a base structure to identify the scope and extent of the obligations of the data controller in the GDPR. This ontology has been divided into three main areas:

- data protection principles;
- rules of data processing;
- data subject rights.

Although the principles are general concepts that evolved in the last decades, the rules of data processing and the rights of the data subject are provisions of the regulation. The rules represent direct duties to the data controller, whereas the data subject rights can be seen as complementary to duties of the data controller: as [20] point out when building their ontology, where symmetric roles are assumed when mapping data controller's duties to data subject's rights.

On the other hand, [19] identify the controller as responsible for implementing data protection policies and security measures: during process development (privacy by design), as well as in terms of default user settings (privacy by default).

Data protection by design use case requirements include implementing PETs, encryption during storage and transfer, access control, pseudonymization, among others.

Data protection by default use case requirements include data subject explicit consent over the processing of his data, transparency, data protection impact assessment, data minimization, collection of data for a specific purpose, among others.

Both [19, 20] allow for a good understanding of how to organize the content of data regulation, but do not give insight on how to translate that knowledge to real implementations. Thereby, there is a need to represent data protection concepts so they can be modeled with system processes.

Not so focused on the concepts of role and duty, [24] starts by inferring the foundational concepts involved in protecting privacy. Four foundational concepts were identified, plus three support concepts:

Data Different levels of data categories such as: Personal identifier information; Sensitive information (credit card numbers, medical data); Medium sensitivity information (age, gender, birth date, and social relations); Usage data (browsing and transactions history); Public data (any data that is made public by users).

User Set of individuals or organizations who can access the data, distinct from a privacy point of view: Subject; Provider; Controller; Processor; Third party; Recipient

Purpose Describe for what purposes the data is collected, used, or disclosed. There are several types of purposes, those that are related to this research are: Completing a task; Analysis; Marketing.

Action Activities performed on data (Collect, Read, Update, Disclose, Delete).

Permissions Granted to users to access data (Permitted, Forbidden, Permitted if condition is true) and linked to actions, where an action is performed only if it has the permission to do so. Also, permission cannot be given unless a purpose is specified.

Conditions Conditions that must be true to allow an action to be performed on data.

Data Retention Defines the period of time the data is kept at the requester end.

Using the above described concepts, [24] specified the following requirements that were then used to develop the privacy annotations as their BPMN extension(later in this text): Access control; Separation of tasks; Binding of tasks; Necessity to know; and User consent. Although the first four requirements were based on previous studies [25–27], the *User consent* was included due to its importance in preserving user privacy and making sure that the user trusts that the system will not used the data without consent. This study aims to a privacy-aware business process modeling and execution framework with support for enforcing privacy constraints during run-time, thus the privacy requirements are modeled on the process level, which requires an extension of the process modeling language BPMN (later in this chapter).

[21] makes a distinction between organizational compliance and process-oriented compliance. Their process model template focus on process-orientated GDPR compliance and contains the GDPR aspect of Data handling and processing, Inquiry of data, Data Breach. The model is designed to be compliant and can be added to existing models. On the other hand organizational compliance describes the behavior of processes applying principles, accountability and governance.

One of the challenges of complying with a regulation is to interpret and translate requirements which are often written in everyday language, using paragraph rules, to strict languages that can be used with an organization system's process descriptions. [22, 28] uses semantic Web technology on specifying and enforcing privacy requirements on access control level. One of their first steps is mapping policies from high-level legal text on Data Protection to operational-level formal languages. The process goes by rewriting these high-level policies using the original natural language, but keeping only the minimal

vocabulary that allows to preserve the meaning and to transform into an if '*condition*' then '*action*' syntax. The vocabulary considered is the following:

- **User categories** *Data controller*, *Data processor* and *Data subject*
- **Data categories** in other words *Personal Data*
- **Data-processing actions** better defining the concept of processing. This goes from collecting, recording, storage, adaptation, alteration, retrieval, use, to disclosure by transmission, dissemination or even just making available. Includes automated and non-automated processing
- **Purposes** legal and legitimate grounds for processing as stated in the European Directive on Data Protection
- **Obligations** specific obligations to the controller or data processor when or after processing personal data

Using the concepts in the considered vocabulary described above they express policies in a simplified way in order to get semi-structured rules. The following structures were presented in [22] where the fields within brackets correspond to the concepts:

A [user category] should be [allowed] the ability to perform [action] on [data category] for [purpose] under [conditions] yielding an obligation to [obligation].

In this case, transforming high-level policies to simple semi-structured rules using an if 'condition' then 'action' syntax would be as the following rule template:

If [Condition on User], [Condition on data], [Condition on Purpose], [Condition on Other] Then Allow [action] and Ask for [Obligation]

The concepts were captured in an OWL[2] ontology where they were defined into OWL classes and properties. The policies were interpreted from text law following the rule template and rewritten as SWRL rules using privacy ontology. The rule conforms syntactically to the SWRL[3] human readable syntax, where an antecedent clause implies a consequence clause. Adapting the rule to an access control policy format, it must conform to the following template:

$$\texttt{Rule} := \texttt{Target} \wedge \texttt{Conditions} \rightarrow \texttt{Effect} \wedge \texttt{Obligations}$$

This template allows SWRL-based privacy policies specifications syntax to conform to the XACML standard[4]. Privacy policies in XACML are composed of a Target, which refers to the resource whose access is being controlled, along with a set of rules. Every Rule contains the specific facts needed for the access control decision-making. It also has an evaluation Effect, which can be either Permit or Deny. The study [28] presents a formalism for mapping SWRL privacy rules into XACML access control rules.

[2]Web Ontology Language [29]
[3]SWRL: A Semantic Web Rule Language Combining OWL and RuleML [30]
[4]eXtensible Access Control Markup Language [31]

The XACML access control model is extended to allow a better control over the contexts with different privacy constraints, when enforcing the privacy policies.

This approach is an attempt to show that the use of Semantic Web technologies can allow both the specification and enforcement of privacy requirements that traditional access control languages and mechanisms cannot achieve. Although important, because it is not developed as a graphical language it can be less accessible to IT employees with lack of information security knowledge.

[32] direct their work to cloud ecosystems where the problems in lack of transparency of data processing are notable, as the data subject is unaware of the risks so he cannot define appropriate policies. The contribution is a solution for privacy enforcement in the cloud, operating on policy specification and generation, data package and transfer, data flow tracking, and policy enforcement.

Their privacy ontology model allows to express data owner preferences as well as regulatory policies into technical policy specification. The PriArmor framework [32] provides automated privacy policy specification translated from the ontology model high-level policies to system call level enforceable ones. The PriArmor agent included in the cloud located PriArmor VM, uses a System Call Interception (SCI) technique to intercept system calls and track all derived data (copies of data, modified data, appended data, etc). The agent then contrasts them with the respective system call level policies allowing the detection of infringements and has the ability to prevent them, providing policy enforcement.

This approach represents a complete solution for cloud data protection enforcement. The simultaneous use of data owner generated privacy polices as well as privacy regulatory policies is a plus. Notwithstanding, it could still improve by using a graphic model notation for expressing the data flow, as managing the translation of high-level policy specification to a system call level is still limited.

2.2 Business Processes

Complying and showing compliance is not an easy task. Data controllers have to adapt their systems to the policies and requirements from the law context where their operate. One of the key difficulties is that data protection revolves around the concept of purpose (as defined in Section 2.1 by [24]) and mainstream programming technologies do not have an obvious concept of purpose [33]. There is, however, one area of computer science where the notion of purpose exists: Business Process Management, in particular Business Process Modeling [33]. Business processes are sequence of related tasks or activities that produces a business goal, service or product. [34, 35]

Organizations collecting and processing data must guarantee that data is actually used for the purpose for which it was collected. This concept has been explored by [33] who questions how to audit a computer system's adherence to a purpose. Their proposal exploits the notion that business process models embodies, by its nature, a particular purpose and also implies when data is collected and used. The main challenge when working business process as purposes under GDPR context is that data transfers between process must be accounted for. The author suggests interpreting process collections as data-flow graphs, introducing the simple model of process collections elucidated in Figure 2.1 where the collections of processes are depicted by which data they use and which data they process and store,

among the relations between them.

Figure 2.1: Data-flow graphs model proposed by [33]

This model is used to develop a methodology for GDPR auditing, where compliance is established if the organization's processes are implemented as described in the process collection models, which describe the system expected behavior. These process models handle data in accordance to privacy policy and conform to GDPR. Also, in turn, the privacy policy must not have statements outside the GDPR.

This methodology provides an interesting point of view when dealing with the need to comply with the right to be forgotten. Data protection regulation impose that data must be deleted on request, provided that the purposes for which it was collected no longer apply. Following the model, where a business process is considered to be equivalent to a purpose, the conditions for erasure are then: no currently running process uses the data and no process that may be started in the future uses the data.

Moreover, this work by Basin et al. stands out by recognizing that GDPR also requires that human activities ensure compliance. The use of business process allows to model human activities besides only automated ones, which is well-suited for the analysis of GDPR compliance as algorithms cannot account for the human activities under the regulation scrutiny.

They only provide the methodology, so even though it partly solves the problem of verifying compliance with GDPR they do not provide an integrated framework which could be deployed into an organization and monitor their processes with easy maintenance.

2.2.1 Graphical Modeling Languages

Graphical modeling languages are very useful because they can describe business process in an understandable way for both humans and computer systems – with these graphical tools it is easier to make rules and to design systems in a way that enables monitoring the compliance to the regulation on data protection, that affects companies and their operations. The following works try to make this connection by expressing the law concepts in a more strict and formal but graphical language or model, allowing for the same point of view of security requirements in both IT administrators, security experts or even auditors and lawyers.

From the several studies found, the graphical modeling languages used were UML [19], DCR Graphs [21] and BPMN [24–26, 36–41]. UML was used by [19] in their use case diagrams referred in Section 2.1. The others are described and dissected in the two following chapters.

BPMN

BPMN is a graphical notation, second and latest version published by OMG[5] in 2011 [42]. Was built to specify business processes and to be readily understandable by all users, ranging from business analysts to IT developers. It is an OMG standard with increasing popularity, which means it is already being used by industries and researchers and has a lot of support.

BPMN was created to model business processes and was not designed to handle data privacy requirements, and, as a result, modeling data privacy policies requires some new approaches to the notation or even adding some new structures based on the existing ones. Because the objective is to reinforce business processes with data protection concepts, an exploration of the existing approaches on different uses of the Business Process model and notation is required to a better understanding of what can be done.

BPMN second version has a mechanism that allows extending standard BPMN elements. This mechanism consists of a set of generic extension elements within the meta model that allow the attachment of additional attributes while maintaining full compatibility with the original BPMN core. However, this mechanism, despite providing a well-defined extension interface, lacks a process model for the straightforward development of extensions [43] and no studies were found that implemented security policies and built their extensions to the standard BPMN using this mechanism, with the exception of [24], who extended the model with privacy annotations based on the requirements described on Chapter 2.1.

[26] starts from the existing core elements and corresponding notations in BPMN, in order to create their own Business Process Diagram meta model (This article was published before BPMN 2.0 so it does not take advantage of the new meta model extension classes). After this definition, they extend their meta model with security classes, which contain the security requirements specification. The extension makes use of the BPMN Artifact. This element class can be divided into Data Objects, Groups and Text annotations and was designed to represent additional or specific situations, allowing the addition of marks or indications to the already defined graphical elements. The authors chose to add a padlock symbol with annotations and capital letters, which symbolize (ou represent) each new security requirement. Constraints were associated to each security stereotype specification and specified in OCL[6]. This paper constitutes only a model (meta model) with security requirements that can help to include security considerations into business process analysis and does not apply any enforcement of this requirements in data protection context neither provides for an executable environment.

One approach made the integration of privacy concepts into BPMN using the standard's extension classes making it fully BPMN compliant. [24] provides a framework build based on three ontologies. The BPMN extension is defined by the PrvBPMN Ontology while privacy concepts and domain vocabulary are defined in a Privacy Ontology and Domain Ontology, respectively. The framework pursues two approaches on extending BPMN. The graphical constructs for the user model are symbols, which are added to depict each one of the main requirements (see Section 2.1) and the extended BPMN meta model with SWRL rules. SWRL is used to specify the rules and constraints, allowing support to model

[5]https://www.omg.org/bpmn/
[6]Object Constraints Language [44]

checking and reasoning about privacy requirements. The framework maps the graphical model with symbols to the extended SWRL, which can later, in turn, be mapped into Business Process Execution Language (BPEL)[7] to be executable and ensure compliance of the privacy requirements.

Extending BPMN with annotations leads to limited scalability and limited expressiveness, as only a predefined set of security annotations are possible. In accordance to [40] this can be solved by using BPMN-Query language. This language enables expressing and verifying generic queries over BPMN Model. Their contribution is a framework that allows for modeling in BPMN, which specifies the policies in BPMN-Q and runs queries against the models. The extension to BPMN integrates security requirements to the language. [40] uses a direct approach applying to their BPMN Extension each of the security aspects in the RMIAS reference model as a security requirement: Accountability, Auditability, Authenticity, Availability, Confidentiality, Integrity, Non-Repudiation. In this case, SecBPMN [40] extends BPMN with security annotations, a graphical syntax(the use of symbols), attributes which are linked to an existing BPMN element, for each of the considered security requirements. This extension naturally extends to SecBPMN-Q allowing for policy like queries in the same language, with security capabilities. An algorithm is introduced to check the compliance of the queries and integrated in the BPMN-Q engine.

The current BPMN standard offers two different views on business processes:

- Collaboration diagrams emphasize the order in which tasks of each participant of the process are executed (control-flow centric models).
- Choreography diagrams describe the process from the point of view of the messages that are exchanged among the participants , following a message-flow centric view.

Because there is no model defined in BPMN for modeling from the point of view of data objects, data-flow driven views are difficult to faithfully model with standard BPMN. [36] try to extend BPMN to model artifact-centric processes. The basic building block is an artifact represented in a similar way to Data Objects in the BPMN standard. Artifact-centric models, specify processes from the point of view of the data objects that are manipulated throughout the course of the process. The main idea is to model the object's life cycle, a specification of the artifact's state, using standard BPMN elements, the symbols for tasks, events, and gateways.

BPMN is an imperative model wherein the process model captures all the allowed activity flows and any deviation is invalid. Declarative models take the opposite approach, expressing the process logic of what it must accomplish but allow all activity flows, except if a constraint is set by the system's designer. In other words, imperative describes how the process sequence of activities is exactly, while declarative allows for any sequence of activities provided there are no violation to the constraints. [37] says that is possible to extend BPMN, in a conservative manner, for declarative modeling. His extension, BPMN-D, only extends two basic elements of the standard BPMN: Activity nodes and Sequence flow connectors. Activity nodes may have a set of multiple tasks, and are considered executed when one of them is executed. For more functionally this set can be inclusive or exclusive, in other words the performed activity has to be included in the set, or must not be included in the set of specified tasks. Additionally

[7] BPEL is a language for specifying business process behavior based on Web Services [45]

a free option is available, where an activity can or not be included in the flow. The sequence flow connector elements establish not only a direct order between activities, but also, more ”loose” ordering relationships, as an example, one activity must be after another, but additional, different, activities may be in between. Moreover, the flow connectors allow for activity repetition. These elements have annotations to distinguish themselves from standard ones. BPMN-D can work on top of a standard BPMN engine and provides means of translation to BPMN assuming the set of possible tasks fixed and known.

Alternatives to BPMN

BPMN is not the only graphical notation capable of representing data protection policies. In 2017 a collaboration between researchers at the IT-university of Copenhagen and the danish company Exformatics [21] published a different approach when it comes to formalize the GDPR. By using DCR Graphs [8] they intent to acknowledge if they can clarify the GDPR impact and consequences in companies

Although DCR Graphs notation is also a business process notation and allows for the same modeling of business activities and events as BPMN, DCR Graphs are declarative models. When developing a language for describing data protection concepts, a clear advantage for declarative models is the flexibility that comes with all flow of execution being valid, instead of having to specify a new workflow or rewrite the process for every new requirement as in the imperative modeling. Another key advantage of using a declarative language, is that legal constraints can be formalized in the same notation as the business process, but can be made and verified independently of each other.

DCR graph's activity elements can be added several roles and constraints to provide ordering in the relations between them. The possible relations between activities in DCR graphs are: Condition, Response, Include, Exclude, Milestone, Spawn. The formalization of the requirements regarding consent is explained, as an example of how DCR Graphs with proper constraints could faithfully represent GDPR processes.

The fact that DCR graphs are declarative models means that although order can be established with restrictions, the model allows various paths by definition and doesn't imply that a wrong path is a violation, and with that becomes harder to detect irregularities, making it less appropriated to later transform it in an executable model that can be used to compliance monitoring.

2.3 Privacy Enhancing Technologies (PETs)

Privacy Enhancing Technologies (PETs) are singular technologies that help in the implementation of security measures. However, they work better if properly integrated with each other. The following two approaches try to make it easier for them to work together. The first one with the use of design patterns and the second with the use of an intermediary framework. Data protection by design and by default is being reinforced by the GDPR as one of the biggest responsibilities for a data controller. The first

[8] A theoretical flexible workflow technology invented by Hildebrandt and his former PhD student in: Rao R. Mukkamala at ITU. Mukkamala, Raghava Rao. (2012). A Formal Model For Declarative Workflows Dynamic Condition Response Graphs [46]. Now integrated in a technology solution by Exformatics A/S. Access: `dcrgraphs.com`

consists in the integration of data protection measures in the project as soon as possible, meaning, from the design phase. As to data protection by default, it is directly associated to the data subject's right of having the maximum set of privacy by default, which have to be guaranteed by the controller without being requested by the data owner.

The correct use of PETs is important to minimize the impact of integrating data protection by design and by default principles in data controllers business processes. However, most of the times, the task of best align privacy requirements to the use of the appropriate Privacy Enhancing Technologies (PETs) is not executed with the best expertise or resources as when detailed knowledge of the PETs is not accessible to the developer during both design and implementation stages. [47, 48] propose privacy process patterns in order to create a clear alignment between privacy properties (requirements) and PETs, and encapsulate expert knowledge of PET implementation at the operational level. [47] also integrates these privacy patterns to PriS [49], a privacy-aware system design framework which aims to incorporate privacy requirements into the system design process. The privacy process design is expressed as BPMN process fragments, which allows a structured way for developers to understand how to implement the various privacy concepts, and to identify best pattern for each particular situation. Each pattern is accompanied by a textual description where a definition, the problems that it solves, its benefits and liabilities, and implementation techniques are shown. The business process design pattern is provided for each type of privacy concept. Eight privacy concept types were covered, namely authentication, authorization, anonymity, pseudonymity, unlinkability, undetectability, unobservability and data protection. With such an encapsulation of business process fragments, the authors aim to guide compliance with the system's privacy requirements at the business process level, while at the same time, allowing process patterns to be generic enough to not be dependent on the implementation of a specific privacy-enhancing technology, but rather on the type of the privacy concept.

On a more technical point of view of the problem, PETs may implement distinct types of APIs and use different underlying components leading to incompatibilities which often result in lower levels of privacy as a consequence of the integration of several individual PETs [50]. Data Protection Orchestrator from the WITDOM project [50] solves these issues by applying BPMN to integrate different PETs composing their privacy-preserving capabilities, while taking into account performance requirements. This combination still requires the intervention of privacy experts, but with an eased effort. Although an interesting approach, the goal is to improve privacy by integrating PETs and not to directly monitor personal data.

2.4 Purpose-Based Access Control

Personal Data must be collected for a specific purpose and only processed for that end. Works [51–54] propose privacy protection based on Purpose-based Access control. Role Based Access Control (RBAC) [55] is used as a starting point and is extended to integrate the notion of purpose. [56] presents a purpose control framework with the ultimate goal of auditing data protection compliance.

RBAC models associate individual user to roles and roles to permissions. The user associated with a role is given the authority and responsibilities for that job. Most RBAC models also include a role

hierarchy, a partial order defining a relationship between roles, to facilitate administration tasks [51].

[54] uses the term *Intended Purposes* to describe purposes associated to data. When data is collected, it is assigned, in accordance to the privacy policies, with a purpose which represents the intended use for that data. This purpose must correspond to a valid ground for processing the data, and the reason for accessing said data must correspond to the intended purpose. The term *access purpose* is used to describe the reason a particular process accesses data. When an access to a data item is requested, the access purpose is checked against the intended purposes for the data item. An access is allowed only if the access purpose is included in the implication of the intended purpose. To be able to model privacy policies that may explicitly prohibit access to data for certain purposes, a prohibitive privacy policy is supported in the *intended purposes* concept. This concept consists of two components: *allowed intended purposes* (AIP) and *prohibited intended purposes* (PIP). As pointed out by the author, this can be used to express organization's specific policies. For example, explicitly disallow access to data belonging to some subjects for a particular purpose such as data collected from children under 13, which must not be used for the purpose of *marketing*.

This paper defines three strategies to determine the *access purpose*: it can be stated along with the requests for data access; it can be registered by application ensuring users that used them only access certain data for the associated access purpose; or it can be dynamically determined by the system, based on the current context. The access purpose is validated by the system which grants an access purpose authorization, attributed not to individuals but to roles. However, in the RBAC models the roles are also attributed user permissions, which means managing access purposes authorizations and permissions would become complex. The concept of *conditional role* is introduced as an extended RBAC model, based on role attributes and system attributes. Through *conditional roles* it is possible to describe a specific set of users in a particular system environment to which are given access purpose authorizations.

While the previous concepts were applied to privacy protection in relational database systems, [53] proposed a privacy preserving purpose-access control model based on [54], giving a formal description of the entities and the relationships among these entities, and then used it to specify the requirements for privacy preservation.

The created Purpose-Based Access Control (PBAC) model introduces the classification of data objects into object types allowing to define and administer intended usages and necessary access in terms of object types instead of individual data objects. This allows for more scalable access control solutions as the number of controlled objects grow.

In the PBAC model referred above, the access purposes are declared by the user, and the storage of privacy metadata associated to data adds a large storage overhead. Limitations which are noted by [52], who proposes Dynamic Purpose-based Access Control (DPBAC) model also based in the RBAC model, introducing the separation of access purpose authorization from access decision. The access purpose is now joint with the access permissions, named access purpose permissions. The process of access purpose authorization assigns the tuples access purpose permissions to conditional roles and, enabling a condition role needs dynamic condition evaluation based on subject attributes and system

context, which means dynamically determine the access purpose. Intended purposes are dynamically associated with the requested data during access decision according to privacy policies set by the data subject and kept in *a well-designed hierarchy of private meta-data*, instead of bound to data as labels. During the access decision process, requested data which intended purposes match with the access purpose are sent to the requester.

The aforementioned models aim to protect the privacy of individuals but as [51] explains in their work the data is an important tool for the scientific industry. The increasing privacy protection over collected data usually means data sets lose data quality. Their aim is to preserve privacy of individuals as well as extracting more information. Thus the proposed conditional purpose-based access control model adds to the anterior models the term conditional intended purpose (CIP). Data can be accessed for a particular purpose with some conditions. For example, data subjects may consider that his/her income information can be used for marketing purpose by hiding his/her personal identification information or that his/her income data can be revealed through generalization.

[56] identifies that, until then, there was lack of systematic methods for determining how data was used, making auditing time-consuming and costly, a result from data protection mechanisms which do not provided appropriate support for purpose control. A framework for purpose control is presented for the detection of privacy infringements, composed by three components.

- The data protection policy specifies the access rights: who requested, what actions were requested, and for which purpose (intended purpose).
- Organizational processes describe the business processes and procedures of an organization defining the expected user behavior in order to achieve an organizational goal. COWS [9] is used for the formalization of these processes from BPMN processes.
- Audit trails are a log history of events that represent actions by the users.

The access to data is granted if there exists a data protection statement that matches the access request directly or through a hierarchy. The compliance is determined by whether the audit trail is a valid execution of the organizational processes representing the purposes for which data are meant to be used. Intuitively, if not, the actual data usage is not compliant with the purpose specification.

2.5 Related work contributions and conclusion

From the state of the art review, several proposals suggest solutions that implement data protection requirements and security measures with the intent of preserving individual's data privacy. Most of them based their translation of the law on the data protection principles appointing the requirements they set for controllers with the goal of help them being compliant. Some of them tried to approximate data protection concepts to the concepts used by organization system 's implementation, business processes,

[9] Calculus for Orchestration of Web Services (COWS), a foundational language for service-oriented computing (SOC) whose design has been influenced by WS-BPEL(commonly known as BPEL [45]), the *de facto* standard language for orchestration of web services. [57]

using graphical business process languages. While some try to verify compliance, others also tried to enforce privacy policies. However, there is not yet a standard system, model or framework that allows for data controllers to show they are compliant and assist auditing data protection regulation like the European GDPR. Specifically, it is missing a passive application that can be added to an organization system without redoing all the implementation and that is able to monitor all the personal data related processes and give capability to the organization to perform constant auditing to the GDPR related content, while also being able to show compliance when audits happen. We proposed to solve this by joining together data protection concepts and a business process based intrusion detection system.

Chapter 3

Business Process Intrusion Detection System

The use of an IDS is the logic choice as it already gives the capability of monitoring the system, without the need of creating a framework for the purpose. However, the hard part is that the IDS will have to understand what data must be monitored, and more important, which rules and restriction are applied to the processing of this data. Normally, an IDS concentrates on a low abstraction level monitoring network and system protocols. The processing of personal data consists in a high abstraction level which is usually represented in business process, which allows to model process considering actions over personal and identifying the entities responsible. Naturally, the right choice is an IDS that was built to monitor business processes.

BP-IDS framework is presented in Section 3.1 whereas Section 3.2 explains how business processes are used to monitor a higher abstraction level.

3.1 BP-IDS Framework

BP-IDS is a tool for process monitoring and incident detection in industrial infrastructures equipped with network communication and information technologies. It operates by using multiple passive sensors to collect chunks of information from the network and use them to reconstruct in real-time the executed business process. The reconstructed processes are checked against a specification (baseline) and business rules. Whenever those executed process deviate from the specification, the activity is marked as a possible incident and the infrastructure administrator is notified in real-time. BP-IDS is a distributed app composed by:

- **Engine core:** collects/receives execution information from the several sensors; Holds to the specified baseline so it keeps identifying and validating activities from the chunks of data received; uses the specified network topology so it can discern which sensor belongs to each section and/or device; sends information to the sensors on what to capture; creates alerts for every violation in the

specified business processes or business rules.

- **Several sensors** distributed in the network or target system, with the function to capture and send information to engine:

  - **Network sensor:** passive servers connected in promiscuous mode to a network switch/hub/router; listens to traffic and capture what is requested.

  - **Host sensor:** application installed in the host; capture information from logs.

- **Administration app/interface:** allows the system administrator to specify the baseline and business information as well as the network topology.

- **Monitoring app/interface:** allows analysis of the sequence of events that lead to an alert; forensics function; allows the visualization of the process and other information.

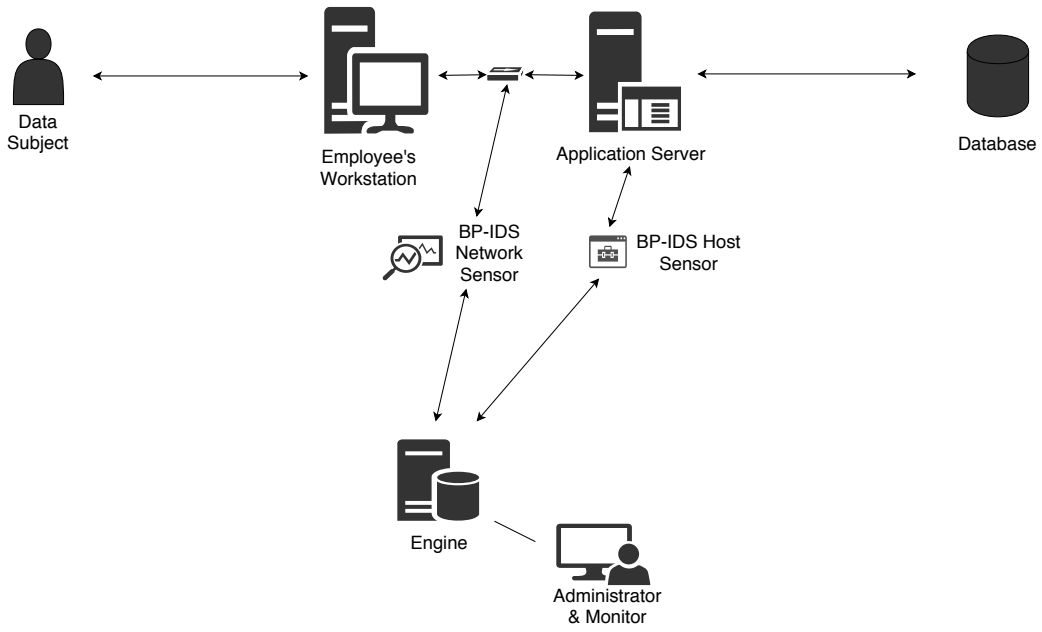

Figure 3.1: BP-IDS system integrated with an application server system

BP-IDS offers a model that allows the representation of real world entities, namely:

- **Informational Entities**

  BP-IDS, besides the representations of the specifications of the business process, defined by a sequence of activities, allows also representations of real world objects. This is implemented by informational entities. An informational entity type must be specified as well as respective attributes, divided into key attributes and non-key attributes. One or more instances of the object can exist by process. Figure 3.2 depicts the general concept for an informational entity type.

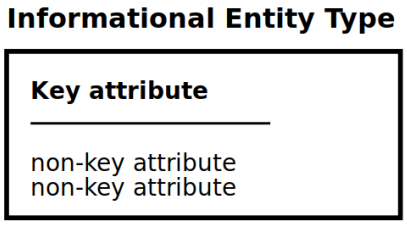

Figure 3.2: Informational Entity Type concept

BP-IDS business rules and validation mechanisms can extend BPMN interpretation giving us a way to express restriction over activities and conditions over informational entities. The tools that give this functionality are:

- **Validator Classes**

  These classes allow to compare attributes from process and from the informational entity associated with the process in verification. Each activity must have one Validator class associated. The classes are part of the BP-IDS engine, and are the fundamental mechanism in maintaining all the attributes of each activity coherent in a process. As an example, if there are two consecutive activities that must have the same value for a determined attribute, there ought to exist a class for each activity that verifies that the next class has the same value (and vice-versa). Algorithm 1 presents the pseudocode for a validator class that gets the attributes from an informational entity and the attributes detected in the current activity, the one that triggered the validator, and compare them. An alert can be send to the monitor by throwing an exception when the conditions for the attributes are not met.

Algorithm 1: Example Validator class

```
Result: Send alert if attributes differ from the process instance
Get process informational entity;
Get detected attributes;
for All Attributes in informational entity do
    if Detected attribute is different from attribute from I.E then
        Create alert description;
        Throw exception with alert description;
    else
        continue;
    end
end
```

- **Business Rules** Business rules allow to maintain restrictions over informational entities' attributes. These rule validations are triggered by changing attributes in informational entities and have to be associated with a rule validator, implemented in a *Java class*. High level rules or restrictions

that affect all informational entities and respective attributes can be implemented. Algorithm 2 corresponds to an example implementation of a rule that is triggered by changes in an informational entity and compares an attribute of that informational entity with the same attribute on all the other informational entities of the same type.

Algorithm 2: Example rule validator class

```
Result: On changing attribute value throw alert based on conditions
Get targeted informational entity;
Get attributes;
conflitIEs is an empty list;
for All informational entities do
    if Attribute equals attribute from current informational entity then
        Create alert description;
        Add currrent IE and description to conflitIEs;
end
Throw exception with conflitIEs list;
```

- **Gateway validator**

  Several Gateways type are possible, among them: Exclusive gateways and parallel gateways; The first requires that only one branch is validated. The right branch is set by a condition in a gateway validator, implemented by a java class. The decision is made based on attributes from the informational entity associated with the process. The parallel gateways do not require that one of the branches is chosen instead of the other. Any of the process branches can be detected next.

3.2 Business Process Monitoring

BP-IDS works by having a business process specification of the system it is monitoring (the baseline). With the several passive sensors it collects chunks of information, from which the engine is able to identify as activities. Every time a chunk of information is collected and verified an event is created. Events can be from several types and each type describes the found activities, attributes, informational entities info and process states. This events are sent to BP-IDS monitor, which is able to reconstruct the sequence of activities that are being monitored, based on the system business process specification.

The process activities are identified from single specific operations in the implementation as for example a SELECT [1] command to the database or even a system call to some file. That means every occurrence of the target resource is detected by the sensors and sent to the BP-IDS engine. In case of infringement, i.e, the activity and respective attributes did not correspond to the baseline, the verification engine generates an event describing the occurrence. The event indicates what is in infringement (failed activity, failed process, business rule validation failed), and contains a description of the occurrence, as well as, associated informational entities.

[1] BP-IDS detects any of the following operations: select, delete, update and insert.

When an event is received and the activity detected, the verification engine verifies the validator classes which were implemented for that activity restricting the use of the attributes. A *failed activity event* is generated in case of any of the values in the attributes do not comply that restrictions. The validator class can also change the values of the attributes.

For the business rules, if an attribute with a rule associated is changed the verification engine verifies the conditions implemented in the rule class and generates an *failed business rule validation event* in case of infringements, likewise to what happens for the validator class.

Identification Patterns

So that an activity can be recognized in a log, an identification pattern must be set. This pattern corresponds to an unique line in the log that matches an activity. Each activity must have an unique identifiable line in the log file. The same applies to components intercepting the database operations that read or write over data in the database. Identifications patterns must be configured for the identification of resources of interest from specific tables.

Collect Patterns

For each operation of interest identified by the previous patterns, collect patterns are set o collect the parameters in the log line correspondent to personal data and related information. These are collected to fill the correspond attributes of a process or informational entity associated with that process.

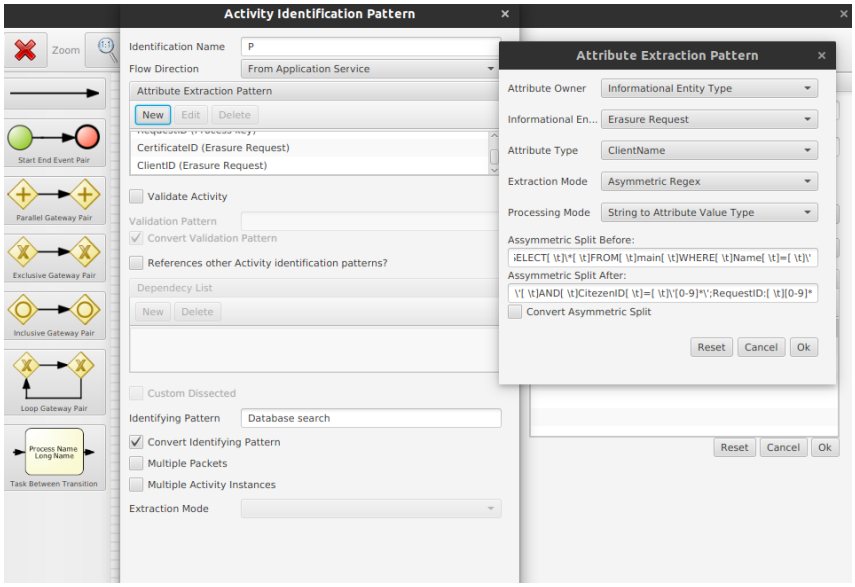

Figure 3.3: Identification and collect pattern example

The collector tool, in Figure 3.3, uses regular expressions to identify the attributes that it needs from the log lines. The expressions have to be manually created and inserted for each attribute.

Chapter 4

Data Protection Compliance Monitoring Solution

This thesis proposes to use BP-IDS, a Business Process based Intrusion Detection System, for compliance monitoring in data protection context. The fact that GDPR requires data controllers to be able to show they are compliant, while maintaining transparency to data subjects, creates a new urgency for data protection compliance monitoring solutions. The challenge can be divided into three phases.

First, this work identifies the data protection requirements an organization must comply with. It also identifies the concepts in the data protection domain that need to be monitored in order to verify the compliance of the requirements.

Second, actions over personal data will be defined from business processes in such a manner that BPMN specifications concerning the personal data flow can be modeled. This specification needs to assure the data protection requirements identified in the first phase.

Last, install BP-IDS into an organization system by setting up sensors and configurations. The selected specification establishes a baseline for the IDS to work and assures BP-IDS can monitor data protection requirements and check for non-compliant behavior in the system, that does not correspond to the created baseline model.

The remainder of this chapter presents how to use Business Process Intrusion Detection system to monitor data protection requirements. Section 4.1 describes how to enforce data protection requirements into business processes dissecting principles, data subject rights, regulation and the integration into privacy policies. Furthermore, Section 4.2 will describe how to integrate BP-IDS with an organization system using BPMN specifications of the processes and configurations derived from the data privacy policy. This section also develops on how violations to the requirements can be monitored with the solution presented and known limitations.

4.1 Data protection compliance

To accomplish the objective of a compliance monitoring application based on BP-IDS, GDPR is used to identify entities, concepts and requirements.

Starting with the various entities that establish the domain of data protection in Section 4.1.1. After that, in Section 4.1.2, data controller duties and data subjects rights are analyzed and the requirements for organizations delineated from them.

4.1.1 Entities

In the data protection context it is fundamental to locate and follow data flows while identifying who accesses data and why. To be able to find the critical data flow one must identify some main elements in a system or distributed system responsible for data processing. We identify the main entities taking in account the GDPR, which defines the following among others:

Personal Data In accordance to GDPR, data must be collected for a purpose, processed for that purpose only and erased if that purpose is no longer valid. This entity will be closely related to the concept of purpose. Data is owned by the data subject and processed by the data controller and his partners.

Data Subject On the data protection context the privacy of the data subject is the main of interest, which puts the data subject as necessary as the source of data and privacy preferences. When the data subject gives his data to the controller, he has to consent to the processing of his data. Exceptions apply when the data controller has a legal ground for processing the data (contracts, legitimate interests, etc.).

Data Controller is the entity that collects data and determines the means of the processing. It is the entity responsible to manage the purposes associated to the collected data.

Data Processor is the entity that processes data in behalf of the controller and respecting the privacy policies set by the controller. Data processor does not control the data and cannot change the purpose or use of the particular set of data.

Data authority Independent public authorities responsible for audit the application of the data protection regulations.

The complexity can be reduced by assuming the essential entities and their interactions as depicted in Figure 4.1. From the data protection point of view the system has to revolve around the concept of personal data. The flow of data goes from the data subject to the controller and then to third party processors if it is the case.

Although this can be a complex system with several controllers (groups of undertakings) and several third party processors the privacy policies applied must follow the ones presented to the data subject when personal data is collected. The controller responsible for the data collection must guarantee all his

partners follow the same privacy policies and are compliant to the same degree [16, article 28]. When developing this solution the relationships in Figure 4.1 will be assumed. The data controller represents the entity responsible for controlling and processing personal data.

Figure 4.1: Essential relations between data protection domain entities

4.1.2 Data protection requirements

Requirements

In order for a process to be validated as compliant with data protection regulations, the participating actors were identified. A description of their roles is now required. In data protection context the ones of interest are the duties of the data controller. GDPR declares in [16, article 24] the responsibilities of the controller as the following:

> *"Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation."*

In accordance with the data protection ontology created by [20], the principles are taken as foundations and the rights and rules are considered applications of the principles. Furthermore, the duties of the controller are considered to be imposed by the principles, rules provided in the data protection regulation and the responsibility of assuring the data subject rights.

The controller and processor have the responsibility to always assure every single principle in Section 1.2.2, which represents requirements that must be guaranteed (i.e., the duties of the controller). Following what was said in Chapter 1, data protection principles aim giving the data subject privacy control over his data. To accomplish this, the principles of data protection give primacy to the concept of purpose of processing. There is an emphasis on better control over the purpose of data collection and the purpose for which data is processed by the controller itself and third party processors that may have access to data. Having a purpose means data is collected and processed with a valid legal ground, justified by legal or legitimate interests or explicit consent from the data subject. The former affirmation can be rephrased to: the data controller can only collect and process personal data with a valid purpose. In accordance with the *Purpose limitation* principle, the controller has (the responsibility) to respect the purpose when processing personal data. Additionally, from the *Purpose Limitation* also results that an access control must be implemented to guarantee data is only accessed for the right purpose. *Data minimization* principle obliges the controller to keep the collecting limited to what is absolutely necessary and relevant to the processing. Besides, the controller or processor who keeps data has to guarantee

data is accurate, requiring regular updates if it is still needed for the collected purpose. For the remaining of the principles the same direct responsibilities can be formalized.

Data subject rights as requirements

In addition to the aforementioned requirements, the data controller must ensure that the processing allows the data subject to exercise the rights to which he or she is entitled by the law. As said in the beginning of this section, data subject rights are applications of the principles; therefore they impose similar responsibilities to the data controller. Besides, it is a GDPR requirement to ensure that the data subject is able to benefit from his rights. As an example to this the Right to Access is considered. In accordance with [16, art. 15], the data subject has the right to obtain from the controller, if personal data concerning him/her is being processed, access to that same personal data and all the information related to it. Following the same line of thought as before, this represents a direct responsibility of the controller in giving to the data subject this information when asked. Although this right is an application to the first principle presented, *transparency principle*, the responsibility to answer to this right can lead to extra requirements for the data controller, which in this case may be the need for means for the data being requested and delivered. The remaining of the rights are similarly structured, the controller must respect them and implement the required means and extra requirements to guarantee the data subject rights.

Direct GDPR requirements

GDPR also includes direct provisions that bind the controller. Some examples are third party processing rules, risk assessment obligations, DPO assignment to some cases and joint controllers rules, which may represent requirements to the data controller.

Privacy policies

Due to the fact that nowadays data processing is mostly done by computer systems, the implementation of requirements applies to most organizations. GDPR also imposes that organizations apply privacy by design and by default meaning all the above principles must be followed and taken in account in all system implementations and structures and should not be a reaction to a request by the data subject. The data controller must implement technology when required (e.g. backup, encryption, etc.). For the requirements that are translated from principles and rights, the controller should model all the requirements to which it is subject to into its privacy policies.

When processing data the controller must have a privacy policy, which decides the behavior of the organization as a whole when handling data. Behavior which, must be compliant with the regulation making the verification that the data processing follows the privacy polices essential when monitoring compliance. In this work we are defining privacy policies as following:

Privacy policy is a collection of statements that apply rules over all data or restrictions over business processes and may be an application of the code of conduct in force in the organization, if there is

one.

When applying data protection requirements into the privacy policy there are two options:

- Rules that are applied directly to an action over personal data, thus statements as they are in the privacy policy. For example, the purpose limitation principle can generate a policy rule which prevents data from being accessed with a purpose for which it was not collected;
- Conditions over specific procedures in the processing of personal data that can be modeled into business process, were activities are restricted or modified to oblige to a statement. A restriction over the activities that handle data can be for example an access to data in a process cannot happen after the activity that modifies data. Conditions over activities sequence should transferred into statements in the privacy policy, but can be implemented by business process graphical languages (e.g. BPMN) extended to support privacy restrictions.

Concepts

To assure compliance in a system that handles personal data, one must be able to model data protection features into the specification of the system business processes. Influenced by the work of [24] and [33] the following four concepts will be defined so to describe data protection requirements over business processes:

Data Groups/Collection of data When talking about data it is not possible to model or monitor all data, besides it would be an insane task to have a different purpose for each individual data. For that, collections of data or data groups are defined. Each group is made from data collected and processed for a common purpose and the same restrictions.

Purpose Representation of a valid legal ground for collecting and processing a collection of data.

Rules Privacy policies applied over all data, applies to all data without context. An example would be: a collection of data can only exist if a valid purpose is affiliated.

Restrictions Process specific conditions applied to data under the process context. An example would be: an already deleted collection of data cannot be access again.

Figure 4.2 illustrates the relation between these concepts, where the solid square boxes represent entities intervening, and in ellipse shaped the concepts described above.

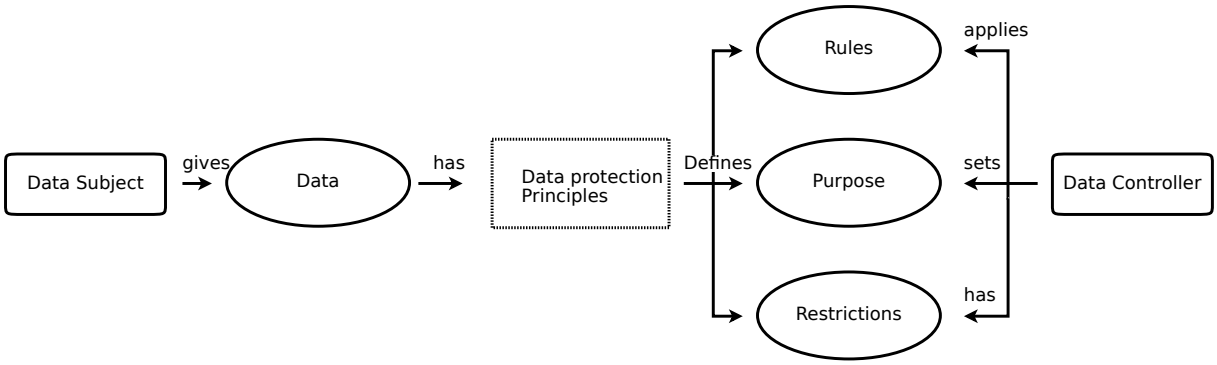

Figure 4.2: Relations between data protection principles and Entities

4.2 Compliance monitoring with BP-IDS

This section, a full guide is provided on how to monitor data protection compliance on a generic target application server system type. The first challenge is to specify the processes that handle data into BPMN, in such a way that they can be read by BP-IDS. The point when doing this is that personal data is the only concern, thus the specified processes must be a simple representation of the personal data flow between entities. The second challenge is to be able to identify the data protection concepts that characterize these processes in such a way that allows us to keep track of them. This should be accomplished with BP-IDS framework.

For the rest of this section a more detailed explanation is provided, starting by the introduction of a generic hypothetical target system on the Section 4.2.1. Then a description, in general instructions, of the physical installation, and after, a guide through the entire setup in Section 4.2.2. Finally, Section 4.2.3, further clarifies how this solution accomplishes compliance monitoring.

4.2.1 Target system type

A general application server usually presents a distributed system network with the architecture in Figure 4.3:

- **Server:** hosts several applications corresponding to each department or function.
- **Database:** The database stores all data and documents from the system. Several apps can have access to it. Multiple or one database can exist, depending on the structure desired. The access to the database server can only be successful through the application server
- **Client:** The intermediary between the service and the data subject. An example of this is the employee receiving data from the data subject, or a browser instance receiving data directly from the data subject.

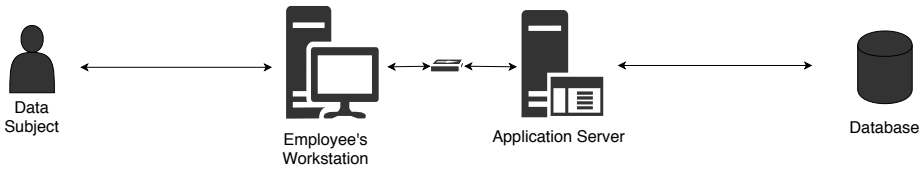

Figure 4.3: Generic target type system architecture

The flow of personal data goes from the source (data subject) to the database, and per request, it may be accessed through the application. The employee has access to data collections when executing a service with access to the data. Table 4.1 depicts a possible structure of the data in a database table. An internal ID is associated with a data subject (tableID) and data is associated with that ID.

tableid	type
id	integer
name	string
age	int

Table 4.1: Generic Application Server Table

As an application based system, the server provides the bridge between the data and the users responsible for executing services. It is assumed that a normal application process or business process always follow the same generic template. A process is started by an event which is normally a request for a service and then an access to the database for information gathering. Data processing may occur as well as further databases access. The service provides a result, which may be returned to the data subject or not. This represent an over simplistic model, depicted in the simplified process example in Figure 4.4, that allows the explanation and validation of this thesis.

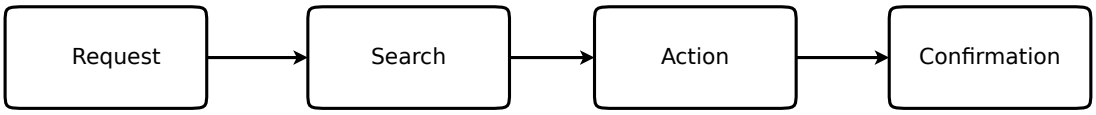

Figure 4.4: Generic Application process

4.2.2 Setup

Physical Installation

The installation requires 3 steps: the sensors must be configured in the application servers and if needed in the organization network equipment (i.e. switches); BP-IDS engine must be installed in one physical or virtual server; a database server must configured for BP-IDS. The sensors must be able to communicate through the network with the engine. The Administrator and Monitor applications must be able to communicate with the database, while accessible to the final user. In practice, the result should be a topology like the one in Figure 4.5.

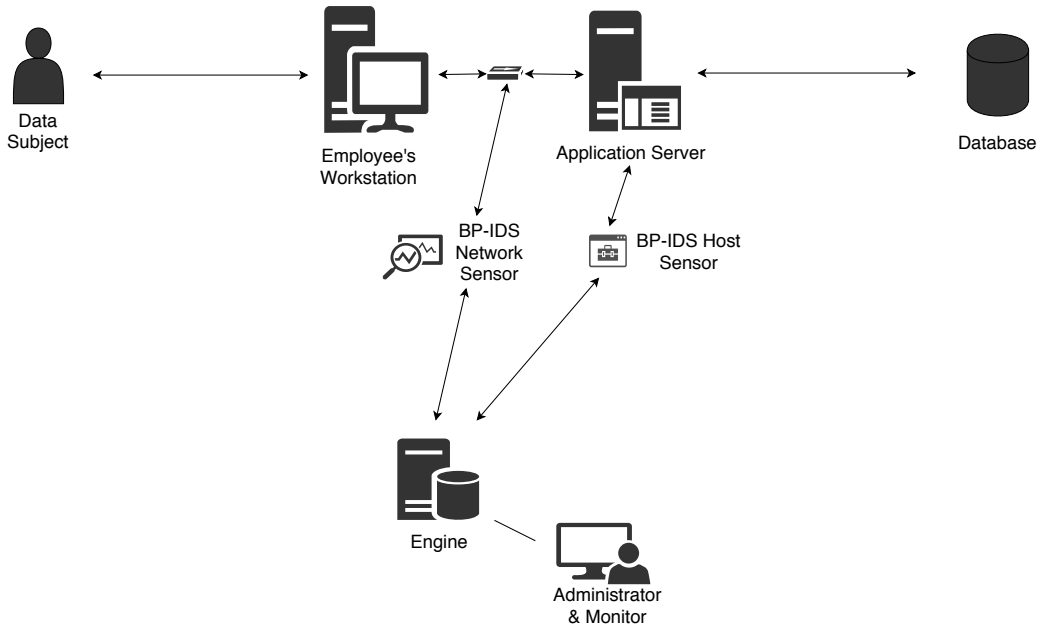

Figure 4.5: BP-IDS integrated with the application server system

Configuration

The organization processes must be modeled using BPMN and the data protection concepts configured in BP-IDS administrator tool, which includes setting up the rules and restrictions from the privacy policies. This setup includes the following 4 steps: design BPMN processes to specify the processes of the organization and the information systems that support them, configure informational entities to introduce data protection concepts, implement activity validators to create restriction over processes, and lastly implement business rules to create general rules over all data. These steps will now be further explained:

1. **BPMN process specification**

   The first step is to model a BPMN specification of the organization's personal data related processes. The organization system administrator and the DPO or other employee accountable for the GDPR should specify the system processes into BPMN. This specifications must be a simple representation of the complex process, as the intention is not to verify all the complex decisions in a system, but just the processing of personal data. Moreover, they should represent the expected behavior of the system and user actions over personal data and must be fully compliant with the data protection policy in force in the organization. These are the BP-IDS baseline processes that allow to detect infringements to the data protection requirements. Any undocumented missing processes may trigger false positives, in such a sense that activities which are not in the baseline specifications may be detected when accessing data.

   Although the aforementioned step can require a considerable effort, under the GDPR, the organization should know where all personal data is stored and handled. The monitoring of their flow of data is an important progress in the direction of a GDPR compliant system, as it allows an organization to track the personal data in their system. For example, if the data from a certain table in the

organization database corresponds to a data group and every access to that table is monitored, any accesses to that data that were not specified in the original baseline specification will be detected. In that way, it allows BP-IDS to detect a misuse of data and represents an opportunity for the organization to correct their system for a more compliant personal data handling.

We want to specify unitary operations over personal data, such as access to data, modify data, or even, add or remove purposes associated with some data (user consented to the processing whereas a contract was already made, giving it additional purposes for the processing). The specification must be sufficient to keep track of all personal data flow. The process should identify the access purpose as well as the entity that is accountable for it. Because these processes isolate accesses to personal data, each process should represent a very particular action over the personal data of an unique data subject, which guarantees that the process can be associated with an identifiable purpose for that action over that data. Furthermore, if each process can identify the entity that executes the operation over the data, then permissions control can be monitored.

Let the generic process in Figure 4.4 represent one application process that receives a request from the data subject, an event; accesses his data; executes some operations; and returns a result which may be a confirmation of an executed service. In this process we only need to monitor the access to personal data (the action over the data) and for what purpose. Figure 4.6 depicts a business process representation of an access to personal data. This access to data BPMN process will be match to the processing of data performed by the generic process just described. The event activity and operation define the purpose for the processing which is assigned to the simple business process. This activity should only be detected when the specified data is accessed for the purpose assigned.

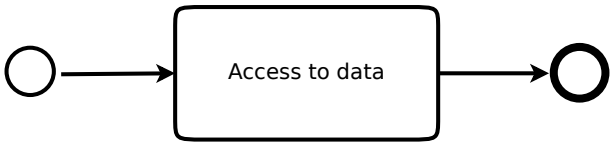

Figure 4.6: Prototype for personal data handling processes - access to data

Even if the target process is rather complex and spans across multiple data groups for multiple purposes it should be divided into simple more practical processes that represent the access to each data group. Figure 4.7 depicts an more complex prototype where a process isolates two actions, the adding or removal of purposes associated with certain data (the modification of the purpose associated with a certain data is considered an action over said data).

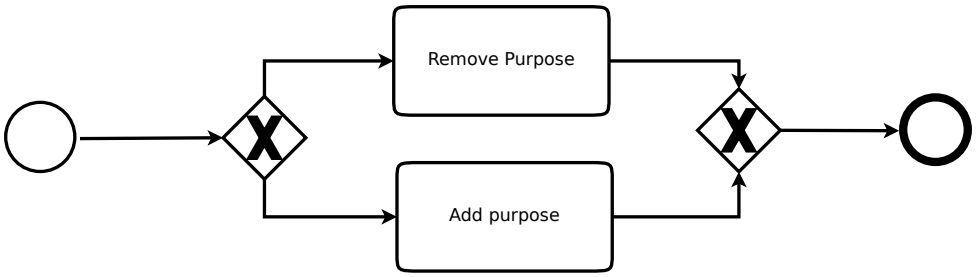

Figure 4.7: Change data purpose prototype example

By using BP-IDS to monitor this process, it will be possible to identify from the fragments picked up in the sensors all the activities completing a process. If one of the activities for that process is not found and the baseline was configured in the BP-IDS with all the activities, a violation has probably occurred. The BPMN diagrams are modeled directly in the administrator user interface of BP-IDS and are kept in the application database.

2. **Informational entities types**

Informational entity type *"personal data"* represents personal data for a data subject. It is identified by the attribute with a data subject ID. The data is not present, because it suffices to know the group which it belongs to and the motives for which it was collected and kept. The informational entity stores the groups of personal data that the data subject has on the system, as well as the purposes for those data groups. But, this only allocates purposes to the data subject. To be valid they must be attributed to data groups. The verification of which data group has each purpose is done by the informational entity *"data-purpose map"*.

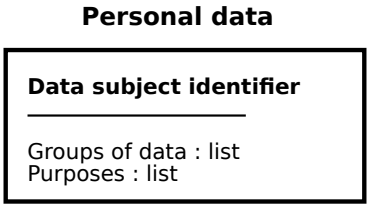

Figure 4.8: Informational entity type personal data

Informational entity type *"data-purpose map"* represents a table that associates one data group to a list of valid purposes assigned to that group. This association should be implemented in the organization system, however, as a compliance monitoring tool the consistency and integrity of this table must be checked. The mapping between data group and the purposes for processing comes from the policies set by the organizations.

3. **Activity validator classes**

The BP-IDS is a passive entity in the GDPR process. This validators act likewise the organization system implementation by following the privacy policy. They are used to model restrictions over BPMN processes. These restrictions can be applications of the requirements asked of the controller or more specific provisions from data protection regulations. Validation is made at attributes

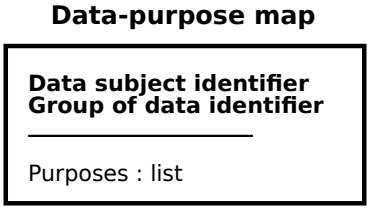

Figure 4.9: Informational entity type data-purpose map

level and with single process scope, which includes the data group or/and purposes used by that process. For each activity to verify the behavior of the attributes, or set restrictions over changes in those attributes, the data administrator must implement a java class. As an example, restrictions can go from validating that the data subject does not change during the process to validating that the data accesses have a valid motive for being used by that process, among others. By considering that a process identifies an entity access, permissions can also be validated, but, although necessary for the GDPR compliance, is not going to be extended in this work.

4. **Business Rules**

   A business rule is set for an informational entity, with the option to specify an attribute or not. The rule is triggered when that informational entity, or attribute if specified, is used. A rule must be associated a rule validator, which corresponds to the logic implementation of the rule, implemented by a *Java class*. This allows to directly define privacy policies, as such as all data must have a valid purpose and all data must be processed for the purpose of collecting. It can be applied over all data or over data groups and has a global scope, not being limited by the process where is triggered.

4.2.3 Application

This solution presents a different approach to the ones presented in the related work. It uses BP-IDS mechanisms and configurations to extend BPMN specification of implementation business processes with data protection policies, derived from the data regulation by the organization, and constantly monitoring the live system implementation.

Although the point is not to make the implementation compliant, in other words, to enforce these privacy policies over the business process, this configuration behaves in very similar way by monitoring all events with the capability to detect a violation to the expected processing. The action which triggers this violation is not suspended or interrupted in any way, endorsing the fact that this is a passive solution. However, an alert is generated with all the information, including the data subject, data group that was accessed, by which process identifying the entity and access purpose, and the purpose associated with the data group. A notification is pushed to the BP-IDS monitor and further actions could be taken, in particular the notification of the data subject and supervisory authorities.

To exemplify how the whole schema works, let consider the hypothetical example in the Figure 4.10, which depicts a simple access to data stored in a database. BP-IDS is configured with the process

Access Data with and the identification patterns are configured to detect activities from the process in the example. When the first activity is detected, BP-IDS creates a new process for access data corresponded to the process type in the example (abstract in this case, should be well defined as it represents part of the *purpose limitation principle* validation). The informational entity for the data subject in question is also created if it did not existed yet. BP-IDS waits to detect the access to the database for that data group. In the informational entity *purpose-map* a relation between the said data groups and the permitted processing purposes should already be filled by a data administrator. Then there is the two step verification to valid the purposes for processing are valid. The process has an associated purpose, related to the type of processing, and the informational entity must have the same purpose in the attributes field of motives. One the other hand, this the purpose must correspond to a relation in the access data group *purpose-map* informational entity. This is verified by the business rule that is triggered by any change to an informational entity. The changes to the informational entity are enforced by an activity validator class in the process activity, which is triggered by the database query for the data group in question. One more clarification: the purpose associated with the process can be defined as a process attribute and must represent the type of the processing executed over data. As an example, if the process accesses data to evaluate the interest of the data subject in a particular brand product to individualized marketing the purpose should be *marketing*.

The example constitutes one way of guaranteeing the *purpose limitation principle*. To be applied to all kinds of operations over data one only needs to configure the activity in the process with an activity validator that triggers the business rule validation.

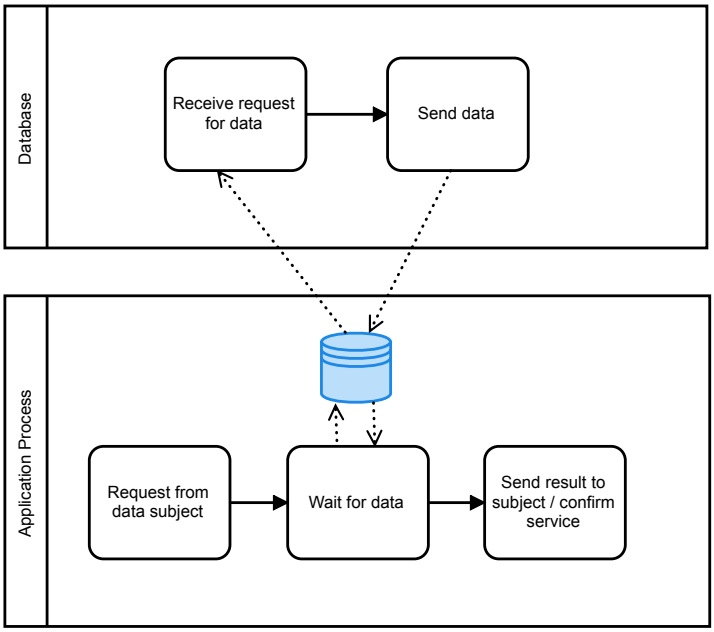

Figure 4.10: BP-IDS example

We should stress out that the basic functionality of the IDS already allows an organization to have a tool that can help auditors and systems administrator to detect any misuse of data by the system that was unknown and not specified in the BPMN specification which represent a violation to the GDPR. By

monitoring accesses to any personal data and trying to associated it with a know process may detect accesses to personal data that were not categorized in the organization's procedures documentation. In this case, if data does not fit to any known group the BP-IDS will generate alert for events of the type *unknown process key*.

The target organization for this solution are small and medium organizations with reasonable simple systems, as for example, town administrative entities. As a passive solution and accessible integration with the host system, it will allow to monitor the processing of personal data and maintain the services it provides compliant with the regulation, giving the users privacy to their data.

Chapter 5

Validation

This chapter validates that a business process based intrusion detection tool can be used to monitor and verify the GDPR requirements when processing personal data. This validation is performed in the context of the COMPACT project, partnering INOV and the Municipality of Amadora. The scenario of the COMPACT trials, implemented in the CMA premises, enabled the evaluation of this solution in a real world system architecture and data.

Adding to the tangible system architecture and data, the solution will be configured to implement an use case for the right to be forgotten. Section 5.1 describes the use case; section 5.2 the architecture; and section 5.3 the necessary setup steps. The evaluation: description of tests and respective results can be found in section 5.4.

5.1 Use Case - The right to be forgotten

Aiming to perform an evaluation to the approach proposed, the developed use case scenario corresponds to one of the changes that GDPR introduces: the right to be forgotten [16, art. 17]. In this use case, the principles of *Purpose limitation* and *Storage Minimization* will be considered.

The right to be forgotten has created some turmoil among organizations when the GDPR entered into force. One of the problems which arose was that just because the data subject chooses to delete his/her data, it does not mean organizations can wipe out all the data. User data is, most of the times, the main resource used for providing services. Moreover, certain data is affected by other regulations, cases where lawful, contractual, or legitimates interests of the organization overpowers the individual right for privacy. In those cases data must be held for a period of time before it can be erased, for as long as several decades. A good example of a case like this are financial transactions records which must be kept for at least 10 years[1]. In fact, according to [58], this requirement ends up not being about forgetting, but deleting outdated data as well as irrelevant data: "*An act of disremembering is impossible to enforce, but the principle has been invoked to support personal privacy by expunging outdated, inaccurate or irrelevant information.*" [58] In the context of the right to be forgotten, two principles stand out: The

[1] This is based on the Portuguese Law (*Código IRC, artigo 123 nº4*) that states that finance accounts must be stored for 10 years

Storage minimization principle which obliges any data, whose collection purposes are no longer valid, to be erased; and *Purpose Limitation* which sets data to be processed only for a valid legal ground. We will be using these to demonstrate that our solution applied to the right to be forgotten is functional.

The fact that personal data can only be processed with a valid purpose and that it is affected by different regulations, reveals the importance of the classification of personal data into collections. Data controllers must divide clearly the different purposes for which data is collected and kept. By doing this, the data controller is able to understand when data may be in a situation of infringement to the regulation; therefore maintaining its system compliant. For example, the classification of the purposes for data collecting by retention period which is the time for legal or law obligations to expire is important to be able to answer to the data subject right to be forgotten request. Data must be associated with the purpose for which it was collected, and furthermore the type of purpose gives the period of time data must be stored. As the example aforementioned, in the case of transaction records where data must be kept for a specific time, a purpose can be attributed to that data: giving it the retention period of the purpose. The association of data groups with a purpose is not unique: allowing data groups to have more than one purpose for being processed.

For the defined use case, the right to be forgotten is interpreted as follows: An organization receives a request from the data subject; It has to search for all data from the data subject and decide which data could be erased. Three options should be possible: the full erasure of all data stored related to that person and associated with identifiable data; a partial erasure of the data that was only collected with consent from the data subject and does not have other legal grounds associated; and finally, no data could be erased due to legal grounds, as an example, data subject's data may be needed for keeping a record. Besides the right to be forgotten request process, the organizations must assure the compliance with the storage limitation principle. They must periodically verify if all personal data on individual subjects still maintains the purposes for which it was collected. In case of expiration of the purposes, the data associated with those must be deleted.

5.2 Architecture for evaluation purposes

The right to be forgotten use case was implemented in the municipality of Amadora premises as described in the remainder of this section.

5.2.1 Infrastructure

The Municipality of Amadora (CMA) is comprised of several departments (such as: financial Department; modernization and technology department; municipal police department, etc.), which work together to offer municipal services ranging from: notary; geographical information; civil defense; and public relations.

The system structure is built upon a central technology infrastructure (CTI), located within CMA premises which offers two functionalities: a dedicated virtual machine (VM), and a central data storage

system accessible by the applications supplied within those VMs. The employees use Remote Desktop Protocol (RDP) to connect to one of the three CTI servers, via the infrastructure's load-balancer. They use the software applications installed in the VM, which will automatically save their documents in the CTI data storages. Figure 5.1 depicts, colored in black: the CTI, consisting of internal servers and SQL server, and the connection to the employee workstation.

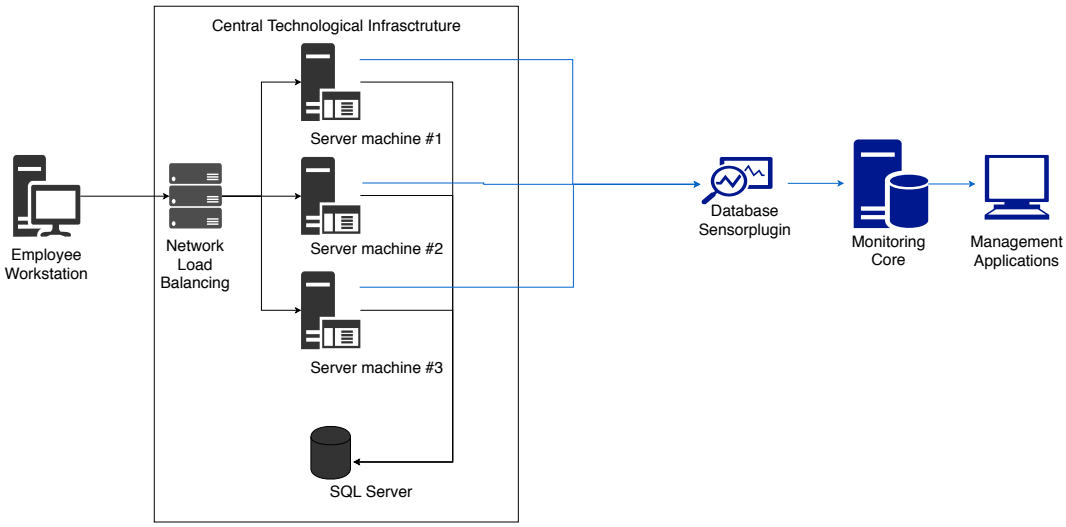

Figure 5.1: BP-IDS installation made on the IT infrastructure of the municipality of Amadora. Coloured in black are interactions that occur in the IT infrastructure, whereas in blue are the additional connections required for BP-IDS to monitor database activity.

5.2.2 Business Processes

In this evaluation, the use case scenario is composed by three sets of processes. The first is a generic department process, that accesses personal data with the purpose of providing a service. In particular, the emission certificate process and the processing of salaries were used as real system examples. The second set of processes resides in a GDPR related process that implements the *right to be forgotten* and the underlying fundamental principles. Lastly, the storage minimization process implements the principle of *storage minimization*. This process is important as it introduces a prevention capability to the solution: by monitoring data which should not be stored anymore due to expired legal ground for processing.

The BP-IDS integration in the target system contributed to mapping out which data tables[2] contain personal data. In total, 34 data tables were monitored. To ease the specification of processes, in this evaluation the data groups were composed by collections of database tables. Each collection of tables represents a data group. The several tables, constituent elements, of each data group were identified to have data with the same retention time from the three options considered for the right to be forgotten in this use case: immediate erasure; erasure after ten years; and never to be deleted. Three table collections were defined, where each is represented by a generic table. The Table 5.1 is indexed by the

[2] *Database server* Tables

table collection names identifying the purpose for data processing, and it is establishing a relation with the retention time: the time that data is expected to have valid legal grounds for processing. Data in the *Employee* table collection is always needed and valid according to GDPR, while the individual is an employee of the organization, and also, after the employee leaves. The same happens with the data belonging to a citizen, stored in the table collection *Citizen*, which is assumed to be kept in the system for at least ten years. Both data table collections may have data that must not be erased or immediately erased. Although, in this context, it is assumed that all data in each collection is equivalent in the data protection context. Finally, the *Other* table collection represents personal data that must be immediately erased when requested (associated with user consent). This may be perceived as an over simplistic model that may not be found in all deployments.

Table collection	**Time to erase**
Employee	never
Citizen	after 10 years
Other	immediate

Table 5.1: Example of the distribution of CMA data tables into groups of data. Each table collection contains data associated to a retention time and to purposes. For evaluation purposes each table represent several real tables in the database containing data classified into the same data group tables.

Generic department service process

The services offered in the municipality system are represented by business processes and rules. However, due to the fact that the aim is to monitor personal data, actions over personal data must be specified. The resulting processes should include the representation of data protection concepts and should be sufficient for BP-IDS to monitor compliance.

The process in Figure 5.2 depicts the major activities in the emission of certificates process, implemented in the organization's system. In this case, one can identify that the process accesses user data, in particular, the necessary information to process both the certificate emission and the respective payment. From this example it is possible to identify the generic process in Figure 4.4; therefore, it could be reduced to an access to data BPMN process. Since the nature of the information may be from several data groups, the access to the data process was designed to differentiate the several tables defined in the use case being followed.

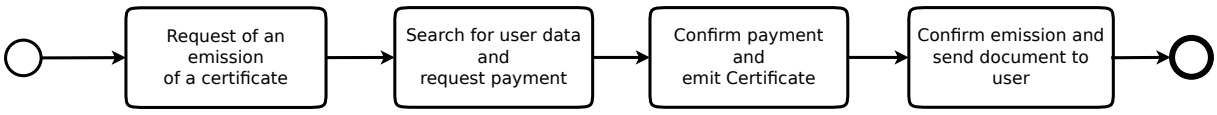

Figure 5.2: Certificate Emission Application process

Following this through, all processes following the same template could be represented by one or several accesses to data. These results led to the process specified in the Figure 5.3. Any process

accessing one of the three tables can be modeled in relation to access over data by this process. The fact that the data groups were fixed with a table collection helps in the creation of these processes and in the validation to confirm it is possible to monitor compliance by isolating accesses to data, when introduced in the BP-IDS tool.

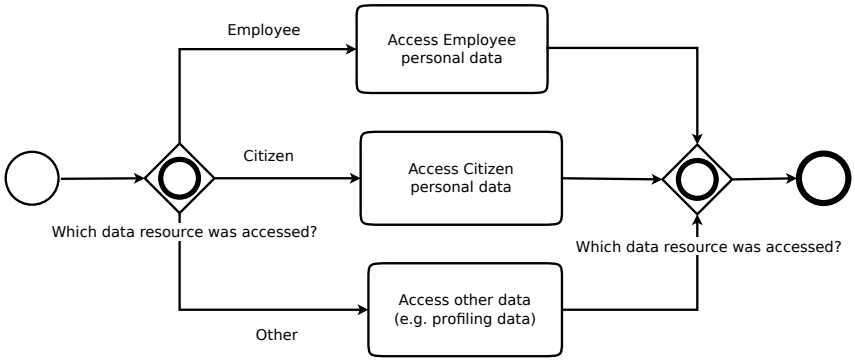

Figure 5.3: Access to data discerned by table

GDPR related processes

At the time of the writing of this document, no specific process existed in the organization to handle the requests and automatically erase the individual's personal data. In order to overcome this, a light software application was developed to implement the GDPR procedures for the trials: form-based application (Appendix A) for managing right-to-be-forgotten requests. This application has three modules (three forms) that simulate three functionalities:

1. allows CMA employees to exercise their right-to-be-forgotten, removing explicit consent to all data stored in the municipality database (A.1);
2. allows the municipality to keep-track of their purposes (legal ground for processing) for storing data (A.2);
3. allows to maintain an association of authorized relations between data groups and valid purposes for processing (A.3).

These forms produce logs in the application server and are monitored by a BP-IDS sensor.

5.2.3 BP-IDS integration

In the context of the COMPACT Project, it has been locally installed on the CMA premises a version of the BP-IDS tool. It was possible to test the monitoring approach of this thesis in the environment configured for the COMPACT trials.

The intrusion detection tool was configured to monitor the interaction between the municipality applications and database, and identify incidents related with the processes in the previously identified: department service applications; and GDPR related data management. To identify data and activities,

BP-IDS was deployed with a sensor [3] capable of identifying business activities in operations made from the servers to the database, whose connection is based on ODBC interfaces. As explained in Chapter 4, BP-IDS is a distributed application that is applied across the monitored infrastructure. In this configuration the following components were installed and are shaped by the additional blue-colored elements in Figure 5.1:

- Several custom tracers (which intercept ODBC operations that read or write over data in the database), installed in each server;
- One sensor that identifies evidences of business activities based on the database interactions intercepted by the tracers;
- One monitoring core that provides incident detection by comparing evidences captured by the sensor with business process specifications;
- One management component containing two types utilities:
  - Administration Interface utility, that allows employees to setup BP-IDS with the necessary specifications;
  - Monitoring Interface utility, that allows employees to follow the monitoring process employed by the BP-IDS tool and conduct forensic investigations to assess the cause and impact the incident had on the organizations devices and business goals.

5.3 BP-IDS Configurations

BP-IDS allows for the implementation of *validator classes*, that allow to verify that the live captured processes correspond to the baseline model and *business rules*, which allow for further constraints over informational entities. By using them it is possible to specify compliant BPMN models of the processes that are going to be evaluated. For the use case being developed, the following configurations were introduced in the BP-IDS Administrator tool:

Processes Diagrams

The following BPMN processes had to be configured in the BP-IDS diagram editor, part of the administrator tool:

- **The right to be forgotten request** was defined as the process that receives a request from the data subject. The key concept in the right to be forgotten is the data subject consent over the processing of his personal data. When the data subject asks to be forgotten, what must happen is that any consent previously given by him in concerns to his data has to be removed. By doing this and by the *purpose limitation principle*, all his personal data that is not being

[3] The sensor (developed by INOV during COMPACT project) is able to identify storage resources accessed and users involved based on information collected from tracing ODBC operations on SQL database connections. More information about the functions monitored by this sensor can be found on: https://docs.microsoft.com/en-us/sql/odbc/reference/syntax/odbc-api-reference?view=sql-server-2017

processed under other legal ground must not be processed any further. Additionally, by the *storage minimization principle*, data without valid purposes must be immediately deleted.

The BPMN specification in Figure 5.4 gives us the organization's implementation of expected behaviour for a request from the user and the decision of what data is to be erased, with the three options earlier defined (Section 5.1). Each of the three branches has an activity validator that removes the consent from the "purposes" field in the associated informational entity.

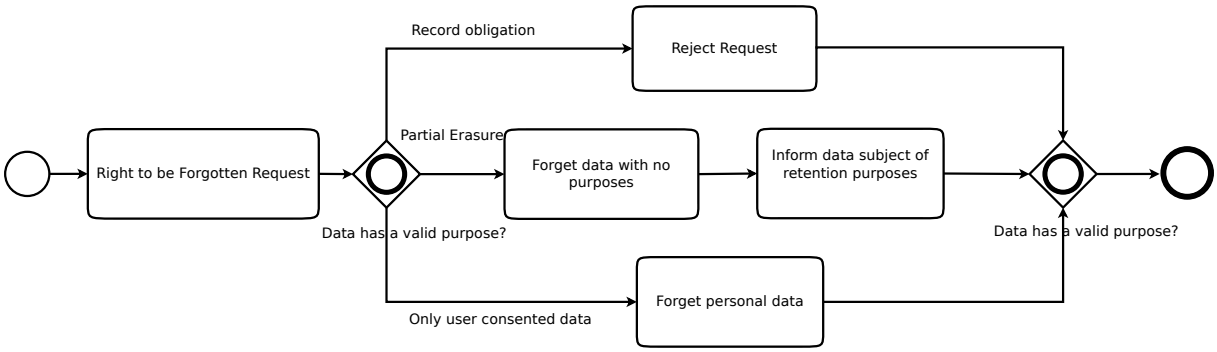

Figure 5.4: Right to be forgotten BPMN process

- **Storage Minimization** Principle validation process is a single activity run periodically by BP-IDS that validates that personal data has valid purposes to be held. The activity validator checks if all data groups for each individual person have valid purposes. This validation process helps prevent inconsistencies by alerting possible infractions to the regulation. The periodic verification of data without valid grounds for being stored raises alarms in situations of non compliance, before an access to said data happens.

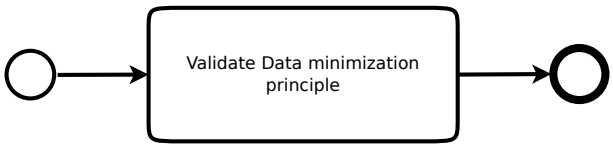

Figure 5.5: Process that validates the verification of the storage minimization principle

- **Access to data** process is a general process that specifies a read or write operation to data. The process should be identified by the entity that generates it and the purpose for access should be identified. The full explanation is in Section 5.2.2.

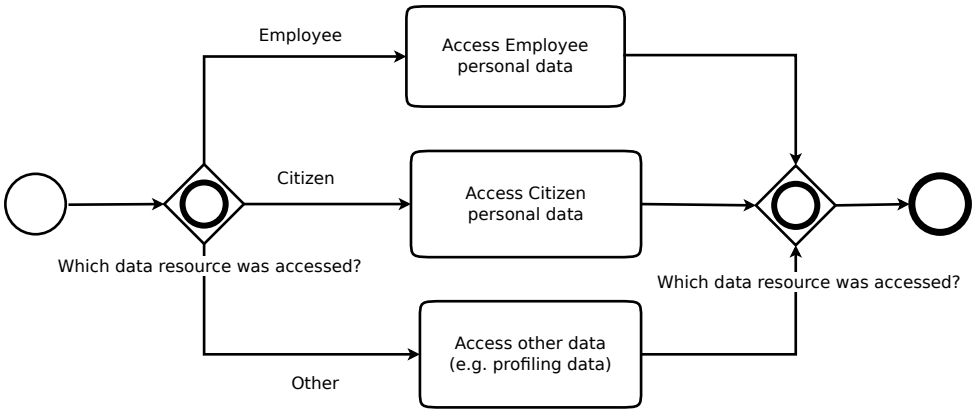

Figure 5.6: Access to data discerned by table

- **Remove Data** process detects removal of data in the host system. Data groups are also removed from the respective informational entity. The creation of this process follows the same approach as the access to data process, isolating the action over personal data. BP-IDS needs to monitor when the data is removed to update the informational entities. This needs to happen for two reasons: to monitor the data that is in a situation of non compliance (without valid purposes); and to avoid false positives, because if BP-IDS has a different information from the host system, then alerts are generated.

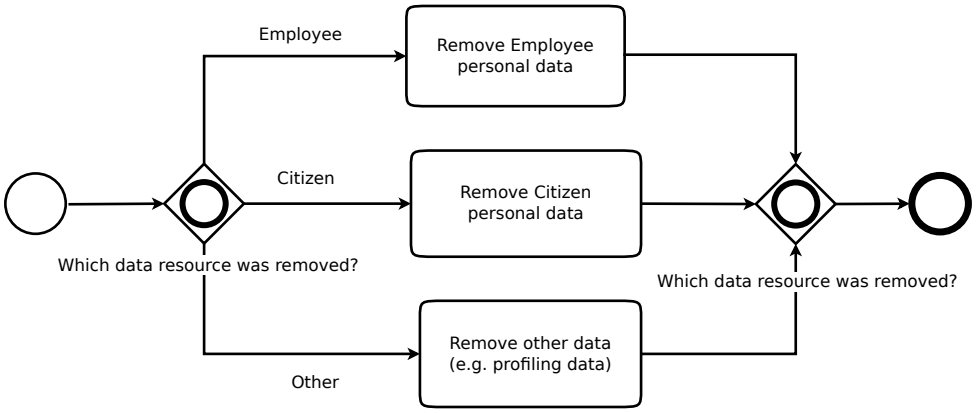

Figure 5.7: Process that detects removal of data from the database

- **Purpose Manager** is a simple BPMN process modeled to detect modifications to the purposes associated with certain personal data. The modifications to the purposes associated to each data group are important for the correct verification of the *storage limitation principle* and *purpose limitation principle* as both depend on the association to a valid purpose. The data should be associated to purposes by the organization who controls it. Notwithstanding, BP-IDS needs to keep the associations to be able to check if data is used for valid purposes.

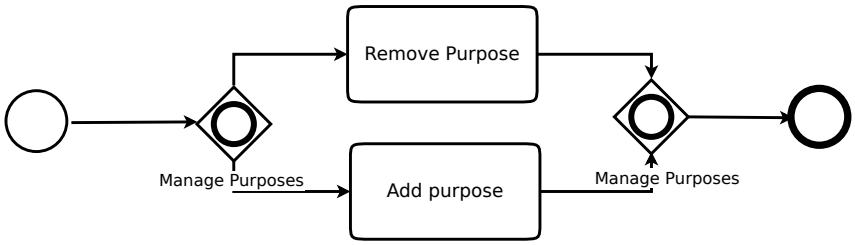

Figure 5.8: Purpose manager BPMN process

Three purposes were considered when testing:

- **Consent** is associated with the data subject consent as a valid legal ground for processing. Data collected for this purpose can be erased immediately.
- **Employee** represents an employment contract where data belonging to the data subject can be used and stored for undetermined time.
- **Citizen** gives purpose for processing data related with a citizen during the provision of services and during a limited period of time afterwards.

Activities validator

BP-IDS activities class validators help specifying processes. They allow, for example, to change attributes values or to validate if an attribute differs from the previous activity. The validator classes that had to be implemented for the business processes created were the following:

- **Add or remove purposes**

  This validator allows BP-IDS to validate the valid purposes for an informational entity. When an activity that changes the purposes to a data group is detected, BP-IDS also removes the purpose from the informational entity.

- **Add or remove groups of data**

  This validator allows BP-IDS to validate if an activity adds or removes groups of data. When data is added to the organization system, new data groups are added to informational entities. The reverse is also monitored.

- **Add data-purpose association**

  This validator allows BP-IDS to update its database with the new data-purpose association. Besides removing the purpose from the informational entity, BP-IDS has to make the relation between the data groups and purposes in the purpose-map informational entity.

- **Validate all personal data**

  Such verification is made by contrasting the attribute list of groups of data to the attribute list of purposes and find out if there is a group of data that does not have a relation in the *data-purpose map*.

Business Rule

A business rule is used for applying one of the requirements directly related to the data protection principles. The verification of the *purpose limitation principle* is directly implemented by a simple rule, triggered by any access to a *personal data* informational entity instance. The rule verifies if all of the informational entity's data groups have a valid purpose.

Informational entities

For testing purposes it was implemented that all data groups may be accessed for the purpose *"Consent"*. Due to the fact that BP-IDS did not have the history of the processing, and thus no past knowledge of data and purpose relations, this purpose is by default associated to all data groups, i.e, all informational entities *Personal data* start with the attribute "purpose" filled with *"Consent"*. In a real environment, all data groups in the system must be associated with a purpose.

5.4 Evaluation

To evaluate the accuracy and trustworthiness of this tool when monitoring compliance, the system was endured with three sets of tests, described in the Sections 5.4.1 to 5.4.3. The first set allows us to validate that data protection concepts could be added to a system BPMN process specification using BP-IDS. The second test set validates that the system can detect compliance deviations caused by attacks and data misuses. Third set defines the applicability and tries to measure accuracy when a violation is detected.

5.4.1 Monitoring Baseline tests

BPMN does not have built-in security constructs. However, as described in Chapter 4, the process specification and the BP-IDS mechanisms can be used together for compliance monitoring. This section aims to validate the thesis question: could a BPMN specification be extended over security concerns using BP-IDS configurations? The requirements to validate are the following:

- the BPMN processes specifications were based on existing documentation and compliant to the regulation;
- the set of processes detects all of the real world data for the defined context;
- the data protection concepts can be monitored, including the monitoring of purposes and implementation of restriction and rules.

Test A – Accesses to personal data

This test focus on generic processes that access data. The aim is to validate that data processing is properly monitored by the detection of the data protection concepts, and that, under a situation with no violations, there are no alerts. Two applications were monitored: the certificated emission application and the processing of salaries application. Both were modeled as described in the previous section. The test consisted in:

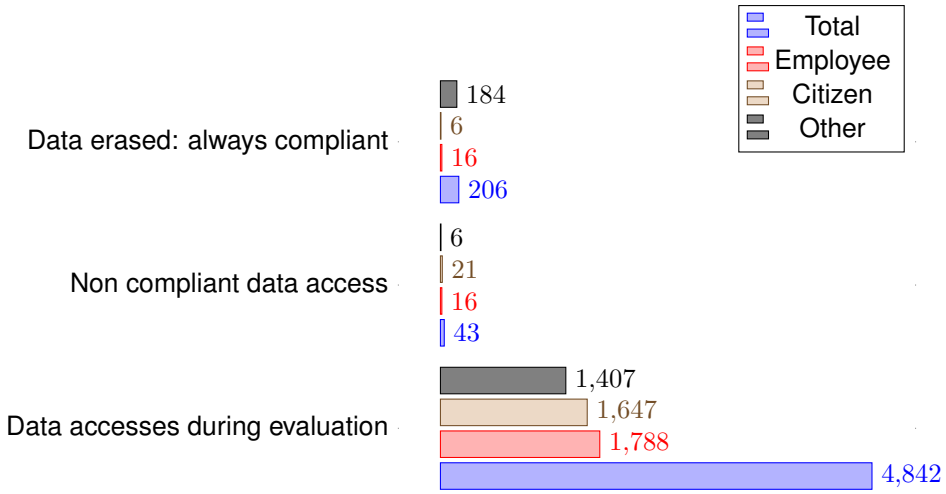

Figure 5.9: Chart with data accesses by table collection

1. Organization's employee uses an application to produce a certificate or to process an employee salary. Because the monitoring is performed on the interface with the database, every different application that accesses the same database is recognized.

Test B – Right to be forgotten request

The aim of this test is to validate if the GDPR related process could be monitored. The execution of specific rules, such as, for example, the removal of purposes from a certain association with a data group, must be detected. This test consists in:

1. Organization's employee fills the Right to be forgotten form (A.1) with the data subject system ID and the data that is to be erased due to not having valid grounds for processing.

Test C – Storage minimization

In this test, an activity validator was configured, in BP-IDS, as described in Section 5.3 to assure the compliance with the *storage minimization principle*.

1. Controller's automated system periodically verifies the system checking data that has no associated purposes.

Could be validated

The specified BPMN processes were based on archived documentation describing the real system, even though only considering accesses to personal data. The monitoring of accesses to data occurs in the interfaces between the application and the database. In fact, what is being monitored are the queries to the database performed by the application. That means that any path of execution the application follows, expected or not, will be detected if a query to the database is performed. That said, we can affirm with a high percentage of certainty that the tests represent a full dataset of possible executions

for the selected processes. Additionally, we can see from the results (Appendix B.1) that no *unknown process key* events were detected during the evaluation period. This event expresses the detection of an activity which is not associated to a process in execution. The absence of those evince that the BPMN specification of the process is correct and there are no problems with the BP-IDS monitoring capability.

The configured setup underwent several tests during the evaluation, which spread over a period of 6 weeks (3 of which coincided with the COMPACT trials). The tool was able to detect accesses to personal data, to right to be forgotten requests and to validations of the storage minimization principle. 5089 processes were monitored, focusing in the personal data of 764 data subjects, 5035 of which were verified and validated as compliant. The remainder 45 activities failed validation and do not comply with the requirements. The following three paragraphs describe the results for each process type.

Access to Data - Figure 5.9 presents statistics for the processes related with accesses to data by table collection (to review the meaning of each table collection see Table 5.1). It can be seen that the non compliant activities represent a small quantity (43) of the total of data accesses (4842). The accesses to personal data are equally distributed across the three table collections. An higher account of data was removed from the *Other* collection. A result that was expected, due to the content of those tables not requiring retention of data for a period time. Table 5.2 represents the sequence of events that occurred during testing on the informational entity 378. The first events correspond to successful accesses to personal data (case A), from: *Citizen* and *Other*. The purpose associated with the informational entity is *consent*, hence the access is valid, as shown by the verified events.

Right to be forgotten - From an attentive analysis of the third and fourth lines from table 5.2, one can observe the data subject has requested to be forgotten (case B). The request was been answered and the data groups associated with the purpose *"Consent"* could be erased, and as a result, the *"Consent"* was been removed from the informational entity (the "purpose" field maintains the purposes associated to that informational entity). However, the data stored for the purpose *"Employee'"* (see Purpose Manager 5.3) could not be erased, which means further accesses to the associated data do not generate alerts. This sequence of events validates it is possible to monitor the execution of a right to be forgotten request. In total, 15 requests to be forgotten were received and answered in accordance to the data protection regulation.

Storage minimization - In Table 5.2, we can see that there is data affected only by the purpose *"Consent"*, which was promptly erased afterwards. If the purpose *"Employee"* is removed, the data becomes available in the system with no purposes to be kept (case C). It was possible to carry out the storage minimization validation: BP-IDS went through all the informational entities in the system at that moment and created alerts when an informational entity data group did not have a valid purpose. The data groups without a valid purposes were marked by BP-IDS as to be erased. Over the evaluation period, this process was run 4 times, 2 of which generated alerts.

The use of BP-IDS could work as an extension to BPMN processes, by adding data protection resources. The next section will focus on the detection of infractions to the compliant processing of personal data.

Process Type	Activity	Failed	Purposes
Access to Data	Citizen, Other		Consent
Manage motives	Add motive		Consent, Employee
Right to be forgotten	Request		Consent, Employee
Right to be forgotten	Forget personal data		Employee
Access to Data	Citizen, Other		Employee
Remove Data	Other		Employee
Manage motives	Remove motive		
Storage minimization		Data stored without purpose	

Table 5.2: Sequence execution from Informational entity number 378

5.4.2 Compliance tests - simulation of violations

In order to evaluate if it is possible to monitor personal data processing across an organization, violations to the expected data processing are simulated. Violations to data protection principles may have internal or external sources and focus on access, modification or storage of data without a valid ground for it. Case A and B present internal abuses or negligence whereas case C describes simulations of external attacks in a structure organized by three levels of data and system access. For each one of them the threat model STRIDE [4] was applied and the most probable threat types tested. For the solution to be valid those case scenarios must validate the following requirements:

- Violations to the business process specification and to the data protection policies must be detected;
- Alerts should be created for anomalies on the processing of data;
- Purpose limitation and storage limitation principles, as well as, the right to be forgotten should be monitored and violations to data protection principles detected.

Case A – Application abuses

In this case, the attacker is the employee of the organization, who wittingly or unwittingly uses the application to process personal data without a valid purpose for it. The application allows access to data without purpose control, which is in non conformity with the regulation.

Access to specific application

- **Information disclosure** Organization's employee executes an application that accesses to a data group that the application should not be accessing (e.g.: The certificate application accesses employee's salary information).

Case B – Application server faults

In the following test, data is not removed and it is kept by the data controller with no valid purposes. Due to implementation design faults, data controller processes do not erase data when they are supposed

[4] STRIDE is a model of threats, used to help reason and find threats to a system [59]

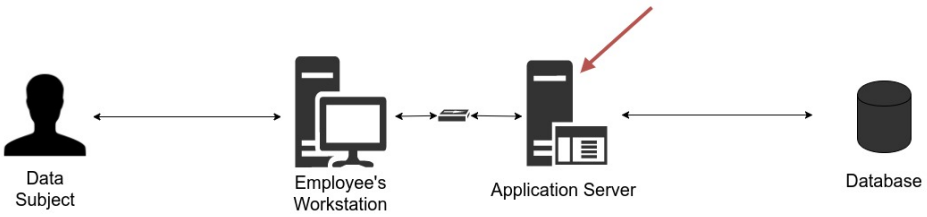

Figure 5.10: Network topology with targeted system indicated by the red arrow - application server

to, leading to the use of data in contexts where the purposes are not valid. The data controller is accountable for this incident.

Server fault

- **Information disclosure** Data subject asks to be forgotten. The employee submits the Right to be Forgotten Form to process the request removing all user consent, however, data is not erased.

Case C – External threats / Cybersecurity incidents

One important requirement for the data controllers is to maintain the integrity and security of the data it keeps. Cyber attacks can have a prejudicial impact on the privacy of data subject if they compromise the data controllers systems. In those cases, data controllers must be able to prove they had everything in order and compliant to the regulation or the consequences may be even worse. Three areas of impact were identified: internal network access and application server access with normal privileges and with administrator privileges. Figures 5.10 and 5.11 present for each case an red arrow indicating the targeted component.

The focus of our approach is to not offer intrusion detection for external attackers; thus, all attackers in this section will be considered to have obtained access to a system login. Therefore, the attacks are abuses of privileges of access over data are going to be considered from either an employee or administrator with access to the application servers and from there the database. In these tests, it will be assumed that the users of the system cannot have direct access to database. The database server is then considered to be off limit and only accessible through the application server.

Access to network

- **Spoofing & elevation of privileges & information disclosure** Attacker accesses data from the outside while spoofing credentials of a certificate emission department employee and accesses the data of a data subject.

Access to application servers - employee/user privileges

- **Tampering & Repudiation & Elevation of Privilege** Attacker modifies the log BP-IDS is using to monitor the host. Attacker modifies the application log to erase previous accesses to unauthorized data.

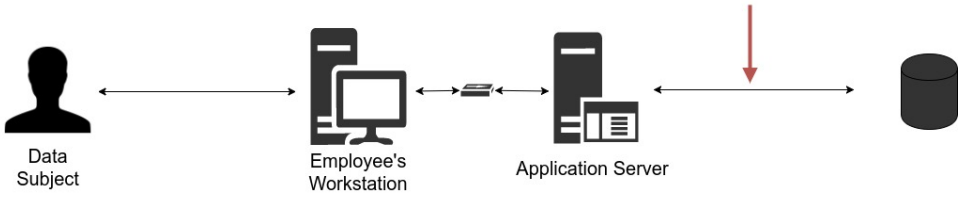

Figure 5.11: Network topology with targeted system indicated by the red arrow - Internal network

Process Type	Activity	Failed	Purposes
Access to Data	Citizen Data		Consent
Manage motives	Add motive		Consent, Employee
Right to be forgotten	Request		Consent, Employee
Right to be forgotten	Forget personal data		Employee
Manage motives	Remove motive		
Access to Data	Citizen Data	Data access without purpose	
Manage motives	Add motive		Consent
Access to Data	Citizen Data		Consent
Right to be forgotten	Request		Consent
Right to be forgotten	Forget personal data		
Access to Data	Citizen Data	Data access without purpose	
Remove Data	Employee, Others		Consent
storage minimization	Citizen Data	Data stored without purpose	Consent

Table 5.3: Sequence execution from Informational entity number 387

Access to database - Admin privileges

- **Information Disclosure & elevation of privileges** Attacker gains access to the system with high privileges and executes a database dump.

Could be validated

According to the collected results (Appendix B.1) , 45 non compliant activities were detected during the evaluation period. From those, 43 were a result from the business rule activation and the other two were related to the storage minimization validation. The three scenario cases above describe the different violations to the data protection that the solution in this dissertation solves. The following paragraphs describe the testing for each case.

Case A - The sequence of testing using informational entity 387 (Table 5.3) had focus in the citizen table collection (which, as explained in Section 5.2.2 represents several tables that contain data required of a citizen of the municipality). Case A violations occur in two situations: when the right to be forgotten request is processed and the purpose *"Consent"* is removed; and when the purposes for processing are removed by an different process in another context, here represented by the *Purpose manager process* (operation performed through the Motive Manager Form A.2). From Table 5.3 we can see that both are detected and an alert is generated: in the second access to data, after the request to be forgotten; and in the fourth access to data, after the *Purpose manager process* has removed the purpose *"Employee"*. These violations were detected by the implemented *business rule* (Section 5.3), that for every access to data verifies if the data group has a valid purpose for the respective process in the informational entity. The third access to the data is legal as the purpose *"Employee"* was added to the informational

entity and the respective relation set in the purpose-map between citizen data group and the purpose *"Employee"*. From the analysis of the monitor tool, 43 alerts were shown, corresponding to the non compliant accesses to data processes. There were no false negatives and also all alerts shown in the monitor tool could be associated to the respective *failed activity event*, which results in a zero percentage of false positives in this evaluation.

Case B - One of the violations of the case B scenario occurs in the end of the event sequence for informational entity 387. All the purposes for processing are removed from the informational entity, which means that all data associated to the person who it represents is in an non-compliant state. The data controller has a limited amount of time to erase it. BP-IDS is able to the detect this "violation" to the storage minimization principle with the periodic business process of the same name. During the evaluation, three infringements were detected: two caused by the execution of the right to be forgotten process (378, 387) and one caused by the removal of purposes by the Purpose Manager process (375). The alerts were correctly generated by the activity validator class, associated with the storage minimization process, and shown in the monitor tool. In this evaluation, all data is initially collected with the purpose *"Consent"*, and for that reason, this validation has failed on two occasions: after the mentioned right to be forgotten; and when the Purpose Manager process was executed to remove all the purposes associated to data. No other informational entity had its purposes removed, and no alerts were generated besides these two, which concludes that no false positives or false negatives were produced.

Case C - Taking into account the threat cases analysed, we could estimate that this solution is able to detect violations to the privacy of the data subject, caused by the external attacks defined in Case C. The isolation of accesses to personal data allows the detection of unauthorized accesses to monitored data. This is possible due to the constant monitoring of data accesses in the interfaces of the application and the database. In fact, it allows to detect any access to the monitored data produced from the application server, while identifying the login of the one producing it. Moreover, once the attacker has a login into the system , he can only access data as an employee (case C, test 1 and test 3)Moreover, the attacker only can interact with the database at the application level; thus is only able to access data as an employee, or, as a system administrator, both of which, are detected as an invalid access to data (case A). Although the testing scenarios were identified, due to the real system environment, it was not possible to the perform further exhaustive tests in this evaluation, as we were only allowed to the applications running in the application server, besides the BP-IDS servers. Thus, it was not possible to access the network or database directly. Therefore, the second test from case C (Access to application servers with employee privileges) was not possible, although it can be seen as an Access to data process violation (case A) for all monitored table collections.

The batch of tests in this section allows to validate that violations to the expected processing of data, as well as to the privacy policies, are detected. In particular: the purpose limitation principle, storage limitation principle and the right to be forgotten are being monitored, along with the respective violations to the requirements. The proposed solution is able to detect a situation of violation and probable infringement to the data regulations and alert the data administrator, even though some limitations were established.

Limitations

- It was not possible to perform attacks in the evaluation scenario. Thus, it is impossible to affirm that an attack would be detected, although the violation to the privacy of the data subject is detected and alerts are created. This solution was designed to identify violations to the privacy of the data subject, but it may not be able to identify if the threat is external to the system. Notwithstanding, an in-depth analysis by the system administrator may be able to detect the source of the attack as processes running and logged in users are available in BP-IDS monitor, even in the cases of external attack.
- We believe that the modification of the application logs result in the lost of integrity of the BP-IDS database, when incorrect or missing information is introduced. This may lead to false positives or false negatives. Although BP-IDS may not be able to validate these activities, *unknown process key* events are generated and the data administrator is informed. Further testing was not possible, as said before, and therefore the impact in the data privacy is not certain. However, this limitation can be tampered-proof by the use of network sensors that complement the information received by the BP-IDS engine. If the reconstruction of a process does not depend uniquely from logs: some resilience is added and two points of trust are created. A detailed analysis from the system administrator could identify the discrepancies in the BP-IDS monitor
- False positives and false negatives are likely to occur from bad specification of processes. Any difference between the BPMN processes specified into BP-IDS and the real system implementation will very likely generate or omit alerts when it should not. The likelihood of this is equal to the likelihood of human errors in the specification of BPMN processes for an entire system. However, for a false positive, an alert is generated and the data admin can easily detect the veracity of the violation by inspecting the events attributes in the BP-IDS monitor. If there is an error in the specification, it can be revised. False negatives can occur from incorrect conditions in activity validators and in business rules, and therefore, it will happen in a similar way for every informational entity that is accessed. A detailed analysis of the BP-IDS logs and BPMN specification is required to identify and revise the specification errors.

5.4.3 Solution behaviour and performance

This last set of test validates the use of this tool for compliance checking production systems. The objectives to validate in this test batch are the following:

- All business activities must be validated within a reasonable delay
- Activities must be validated without human interaction and must not require service interruption to perform security assessment.
- In general, the tool must be fully compatible with the host system
- BP-IDS should not store any personal data collected during monitoring.

Performance

To evaluate the performance, we take in consideration that the GDPR gives data controllers 72h after a breach to inform the supervisory authorities and the data subject. Furthermore, data controllers have 1 month to inform the data subject of actions taken upon his request for exercising his rights.

Figure 5.12: Validation time for each process.

In the chart presented on the Figure 5.12, the interval of time that each process took to validate is depicted. The box plot type chart indicates a medium for each process type and the values that stand out from that medium. It is possible to observe that the validations times are in the order of seconds, even with occasional peak levels associated. The access to data processes indicate occasional isolated values above the medium, while maintaining an average close to zero. This process has an extended range of results, hence it has an accurate medium with its value around zero. For the other two, a minor number of tests where available, and thus a less representative medium is depicted. Nevertheless it is possible to estimate from the values in the chart that for the right to be forgotten process and storage minimization validation process, the validation times would amount to some seconds. The atypical values could have origin as a consequence of network delays or server throttling due to a great amount of events simultaneously. Even if greater network delays are considered, the times for the alerts to be generated are within the limits for data protection.

Applicability

The configuration of the business monitoring services according to the organization's specification is an exhaustive exercise that takes up considerable time and human resources. One needs to know the all organization's monitored business processes and the full extent of the interaction between processes. All process that handle personal data must be accounted for and registered.

On the other hand, the fact that the processes express actions over personal data, it allows for some more generic specification, which is easier to adapt for new processes. By using the same generic department process to specify accesses to data from both the certificate emissions process and the salary payment process, we proved that it is possible to easily adapt or use the specification for identical or similar processes inside the same organization.

The approach of isolating accesses to data can be applied to all organizations whose implementation structure is already described by business processes. Even if it is not the case, data protection regulation as the GDPR may require that organizations have all the procedures in the system well-documented and in archive. If that is the case, the process is simpler in the sense that the processes are already modelled, and only require an adaptation to BPMN.

Moreover, the tool is compatible to system due to its identical architecture, as the one specified in the beginning of this chapter. Besides being compatible with systems using ODBC connectors, network sensors can also be used to collect information which allows the activities to be identified.

Compliance to GDPR

The solution presented in this document is not expected to store the values of individual data. Data is monitored in the form of collections (data groups) classified by purpose and retention period. These collections are associated with an individual person by identifier of the informational entity representative of that person. BP-IDS uses a pseudo-anonymization feature that substituted all personal data featured in the alerts with unique masks that still allowed correlations without disclosing any personal data (since the only way to trace-back the original personal data identifiers is through BP-IDS interfaces). Moreover the actual data does not need to be collected, as it suffices to know the data group to which it belongs. Data belonging to a data groups is subject to the grounds of processing (purposes) assigned to that group; thus, a more grain defined monitoring of the data is not necessary.

Overall, BP-IDS only identifies accesses to data groups and not individually to personal data.

Chapter 6

Conclusion

In the origin of this thesis was a current problem in the context of the data protection in Europe. The new data protection European regulation has set new rules and responsibilities to the data controller. The data controller has now to comply with the regulation, but also needs to show compliance. Moreover, more strict audit rules and penalties for violation the privacy of the data subjects were defined. In need was a solution that allowed to monitor the processing of personal data in an organization and be able to demonstrate compliance as well as prevent possible infringements to the requirements the regulation imposes.

Taking in consideration several studies in data protection ontologies and privacy solutions in the cloud, and also control access based in purposes of action instead of permissions, it was possible to develop a solution to monitor data protection compliance using an business process based intrusion detection system. BP-IDS, an essential infrastructure monitoring tool, is used to help companies verify their compliance and actively check and show they are complying with the data protection regulation in force in their context.

6.1 Thesis hypothesis validation

In the beginning of this thesis, three contribution objectives were formulated. In this section, with the solution implemented and evaluated, it is possible to confirm they are practicable and make sense in the current data protection context.

"*Identify main data protection requirements and concepts that should be monitored for verifying compliance*"

We were able to identify the requirements that one needs to monitor to be able to demonstrate an organization is compliant with a data regulation. To do this we identified the principles in which the data protection regulations are based (emphasis in GDPR). From those the data protection defines data subject rights which we set as application of principles. Direct requirements were also selected from

the data protection regulation provisions. These requirements can be implemented or transferred to the privacy policies which define the behavior of the system when processing data. It is possible to set privacy policies in two formats: rules over action over personal data and conditions over processes.

"*Select the personal data related processes, responsible for personal data flows, and model them as a BPMN specification.*"

We could model processes that represent actions over personal data. The method is to isolate the personal data action in a system process and model BPMN processes that describe the action. Those processes must be simple and represent only an access to determined data. The process must be able to indicate the process that originated the data access, which indicates a purpose for the processing of said data.

"*Assure BP-IDS can monitor personal data related processes and cross check the system current state against the business process specification, following all the rules, detecting any non-compliance*"

By applying the BPMN specification of process to the BP-IDS administrator tool and configuring the activities validator to represent conditions over BPMN processes and BP-IDS business rules to directly impose data protection restriction over personal data we were able to monitor data protection requirements. The BPMN process could be extended to represent the concepts: data, purpose, restrictions and rules. The approach was proved to work in several processes with similar structure without the need for specific adaptation. Violations to data privacy were detected in the testing scenario and information sent to the data administrator for further action.

6.2 Achievements

This dissertation achieved the following contributions:

- A guide to go from regulations to privacy policies and to implementable rules and restriction that BP-IDS can use to monitor data protection compliance in certain environments. The Figure 6.1 helps to visualize the needed steps for being able to monitor data protection compliance in the data processing of an organization. Business process monitoring services can be used in organizations for checking that the implementation of the business process is compliant with its specification.
- Auditing tool for organization to guarantee they are under the data protection regulation, and prevent possible situation of infringement to the local privacy laws.

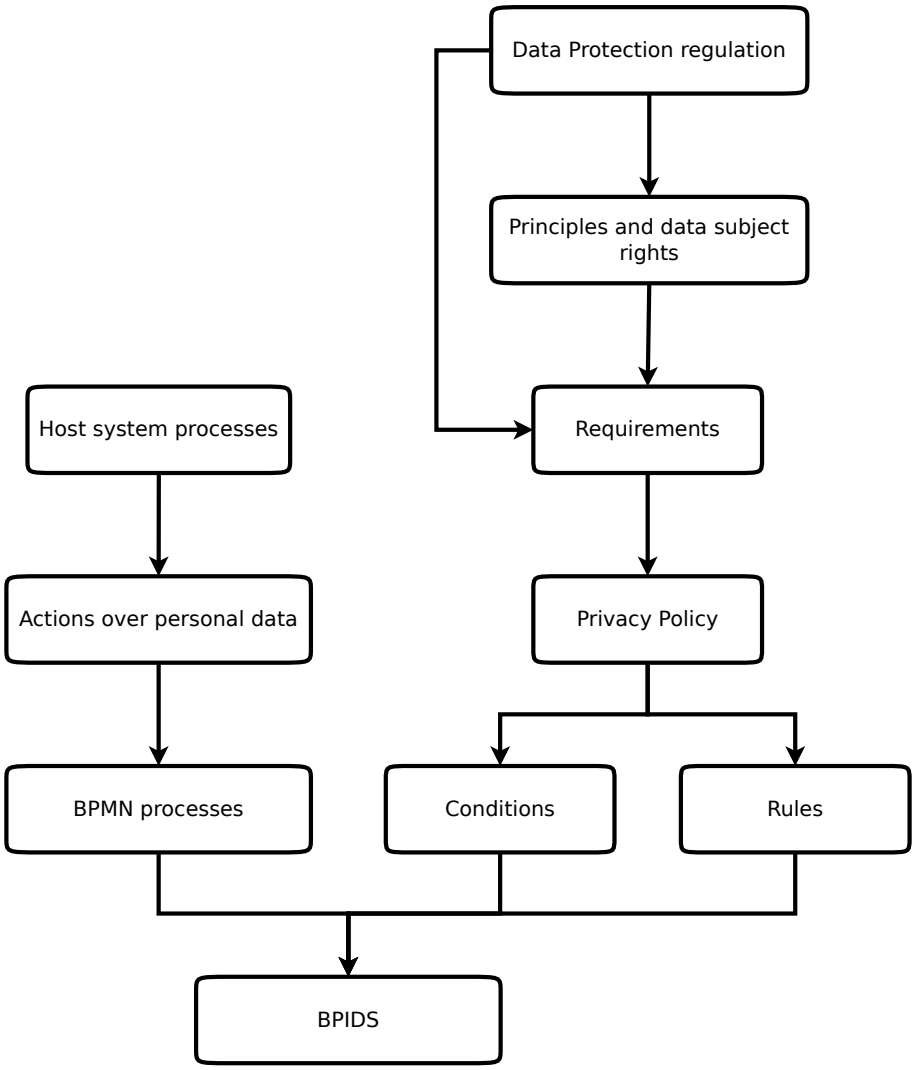

Figure 6.1: Approached followed by this work to implement BP-IDS GDPR solution aiming to monitor the compliance of the personal data processing in an organization

6.3 Future Work

Data monitoring using BP-IDS is still a developing solution and there is still much room for improvement. The paragraphs bellow present some suggestions on where to take from the point the work is left in this thesis:

Specification of the target system Machine learning techniques could be used to automate the search and specification of the target system and the posterior conversion into BPMN processes. This could solve one major limitation by reducing the considerable effort (both in time and human resources) spent by the organization for configuring the business monitoring services according to the organization's specification.

Data protection requirements The approach considered in Figure 6.1 is considered to identify requirements from an data protection regulation. Although we could transcript some requirements to the

testing scenario, the manual configurations of the BP-IDS rules and activities in accordance to the privacy policy is prone to error. An enhanced model should be created to assist the data administrator when mapping the data protection requirements conditions and rules into the BP-IDS processes, activities validator classes and business rules.

Data groups classification The definitions of data groups could be further developed, to better adequate the needs of the real world. The model used, which assigned several tables of data to a hypothetical table and assumed all the data in those table were represented by an equal data group, is insufficient for describing real data. A thoroughly classification of data into categories by the retention time with better description of purposes as legal grounds to data processing requires a more complex model.

Bibliography

[1] K. Granville. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens. URL https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html [Accessed: 2019-01-25].

[2] Handbook on European data protection law.

[3] C. Tankard. What the GDPR means for businesses. 2016(6):5–8. URL http://linkinghub.elsevier.com/retrieve/pii/S1353485816300563 [Accessed: 2018-04-17].

[4] R. Smith, E. Spain, and R. Glancey. *Convention for the Protection of Human Rights and Fundamental Freedoms as Amended by Protocols No. 11 and No. 14*, pages 354–362. Macmillan Education UK. URL http://link.springer.com/10.1007/978-1-137-54504-6_63 [Accessed: 2018-06-15].

[5] D. Banisar and S. Davies. Global Trends in Privacy Protection: An International Survey of Privacy, Data Protection, and Surveillance Laws and Developments, 18 J. Marshall J. Computer & Info. L. 1 (1999). page 113.

[6] E. Council, Forbes Technology. Data Privacy Vs. Data Protection: Understanding The Distinction In Defending Your Data. URL https://www.forbes.com/sites/forbestechcouncil/2018/12/19/data-privacy-vs-data-protection-understanding-the-distinction-in-defending-your-data/ [Accessed: 2019-01-30].

[7] R. Robinson. Data privacy vs. data protection. URL https://blog.ipswitch.com/data-privacy-vs-data-protection [Accessed: 2018-06-15].

[8] B. H. Soares and T. Figueiró. 15 passos para implementar o regulamento geral de proteção de dados. page 20.

[9] Tech-Vendor-Directory-1.4.1-electronic.pdf. URL https://iapp.org/media/pdf/resource_center/Tech-Vendor-Directory-1.4.1-electronic.pdf [Accessed: 2018-03-23].

[10] J. Lima, N. Escravana, and C. Ribeiro. BPIDS-Using Business Model Specification in Intrusion Detection. In *Research in Attacks, Intrusions and Defenses: 17th International Symposium, RAID 2014, Gothenburg, Sweden, September 17-19, 2014, Proceedings*, volume 8688, page 479. Springer.

[11] INOV / BP-IDS – Business Process Intrusion Detection System. URL index.html [Accessed: 2019-05-23].

[12] Home - www.compact-project.eu, . URL https://www.compact-project.eu/en [Accessed: 2019-05-19].

[13] Destaques - INOV INESC Inovação. URL http://www.inov.pt/ [Accessed: 2019-05-23].

[14] Home, . URL http://www.cm-amadora.pt [Accessed: 2019-05-23].

[15] E. COMMISSION. A Digital Single Market Strategy for Europe. URL http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192 [Accessed: 2018-05-08].

[16] 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). page 88.

[17] T. R. Gruber. A translation approach to portable ontology specifications. 5(2):199–220. URL http://linkinghub.elsevier.com/retrieve/pii/S1042814383710083 [Accessed: 2018-03-24].

[18] M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. 11(02):93. URL http://www.journals.cambridge.org/abstract_S0269888900007797 [Accessed: 2018-03-24].

[19] E. Mougiakou and M. Virvou. Based on GDPR privacy in UML: Case of e-learning program. In *2017 8th International Conference on Information, Intelligence, Systems Applications (IISA)*, pages 1–8.

[20] C. Bartolini and R. Muthuri. Reconciling Data Protection Rights and Obligations: An Ontology of the Forthcoming EU Regulation. page 8.

[21] E. Heuck. Digitalising the General Data Protection Regulation with Dynamic Condition Response Graphs. pages 124–134.

[22] H. B. Rahmouni, T. Solomonides, M. C. Mont, and S. Shiu. Privacy compliance and enforcement on European healthgrids: An approach through ontology. 368(1926):4057–4072, . URL http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2010.0169 [Accessed: 2018-05-12].

[23] C. Bartolini, R. Muthuri, and C. Santos. Using Ontologies to Model Data Protection Requirements in Workflows. In M. Otake, S. Kurahashi, Y. Ota, K. Satoh, and D. Bekki, editors, *New Frontiers in Artificial Intelligence*, volume 10091, pages 233–248. Springer International Publishing. URL http://link.springer.com/10.1007/978-3-319-50953-2_17 [Accessed: 2018-03-22].

[24] W. Labda, N. Mehandjiev, and P. Sampaio. Modeling of privacy-aware business processes in BPMN to protect personal data. pages 1399–1405. ACM Press. URL http://dl.acm.org/citation.cfm?doid=2554850.2555014 [Accessed: 2018-03-22].

[25] A. D. Brucker, I. Hang, G. Lückemeyer, and R. Ruparel. SecureBPMN: Modeling and enforcing access control requirements in business processes. page 123. ACM Press. URL http://dl.acm.org/citation.cfm?doid=2295136.2295160 [Accessed: 2018-04-03].

[26] A. Rodriguez, E. Fernandez-Medina, and M. Piattini. A BPMN Extension for the Modeling of Security Requirements in Business Processes. E90-D(4):745–752. URL http://search.ieice.org/bin/summary.php?id=e90-d_4_745&category=D&year=2007&lang=E&abst= [Accessed: 2018-04-04].

[27] Q. He, A. I. Anton, F. Healthcare [hip, G. Leach, and B. Act. *A Framework for Modeling Privacy Requirements in Role Engineering*.

[28] H. B. Rahmouni, K. Munir, M. C. Mont, and T. Solomonides. Semantic Generation of Clouds Privacy Policies. In M. Helfert, F. Desprez, D. Ferguson, F. Leymann, and V. Méndez Munoz, editors, *Cloud Computing and Services Sciences*, volume 512, pages 15–30. Springer International Publishing, . URL http://link.springer.com/10.1007/978-3-319-25414-2_2 [Accessed: 2018-03-22].

[29] OWL Web Ontology Language Reference. URL https://www.w3.org/TR/owl-ref/ [Accessed: 2018-05-14].

[30] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. URL https://www.w3.org/Submission/SWRL/ [Accessed: 2018-05-14].

[31] OASIS eXtensible Access Control Markup Language (XACML) TC — OASIS. URL https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml [Accessed: 2019-05-23].

[32] Z. Primorac, C. Bussoli, and N. Recker. 16th International Scientific Conference on Economic and Social Development – "The Legal Challenges of Modern World". page 855.

[33] D. Basin, S. Debois, and T. Hildebrandt. On Purpose and by Necessity: Compliance under the GDPR. page 18.

[34] M. Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer Science & Business Media.

[35] Business Process Model and Notation (BPMN), Version 2.0. page 538.

[36] N. Lohmann and M. Nyolt. Artifact-Centric Modeling Using BPMN. In G. Pallis, M. Jmaiel, A. Charfi, S. Graupner, Y. Karabulut, S. Guinea, F. Rosenberg, Q. Z. Sheng, C. Pautasso, and S. Ben Mokhtar, editors, *Service-Oriented Computing - ICSOC 2011 Workshops*, volume 7221, pages 54–65. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-642-31875-7_7 [Accessed: 2018-04-05].

[37] G. D. Giacomo, M. Dumas, F. M. Maggi, and M. Montali. Declarative Process Modeling in BPMN. In *Advanced Information Systems Engineering*, Lecture Notes in Computer Science, pages 84–100. Springer, Cham. URL https://link.springer.com/chapter/10.1007/978-3-319-19069-3_6 [Accessed: 2018-04-01].

[38] G. Decker and A. Barros. Interaction Modeling Using BPMN. In A. ter Hofstede, B. Benatallah, and H.-Y. Paik, editors, *Business Process Management Workshops*, volume 4928, pages 208–219. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-540-78238-4_22 [Accessed: 2018-04-12].

[39] G. Decker and M. Weske. Interaction-centric modeling of process choreographies. 36(2):292–312. URL http://linkinghub.elsevier.com/retrieve/pii/S0306437910000591 [Accessed: 2018-04-15].

[40] M. Salnitri, F. Dalpiaz, and P. Giorgini. Modeling and Verifying Security Policies in Business Processes. In I. Bider, K. Gaaloul, J. Krogstie, S. Nurcan, H. A. Proper, R. Schmidt, and P. Soffer, editors, *Enterprise, Business-Process and Information Systems Modeling*, volume 175, pages 200–214. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-662-43745-2_14 [Accessed: 2018-03-23].

[41] A. D. Brucker and I. Hang. Secure and Compliant Implementation of Business Process-Driven Systems. In M. La Rosa and P. Soffer, editors, *Business Process Management Workshops*, volume 132, pages 662–674. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-642-36285-9_66 [Accessed: 2018-04-03].

[42] OMG. Business Process Model and Notation (BPMN), Version 2.0. page 538.

[43] R. Braun and W. Esswein. Classification of Domain-Specific BPMN Extensions. In U. Frank, P. Loucopoulos, O. Pastor, and I. Petrounias, editors, *The Practice of Enterprise Modeling*, volume 197, pages 42–57. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-662-45501-2_4 [Accessed: 2018-07-23].

[44] About the Object Constraint Language Specification Version 2.4. URL https://www.omg.org/spec/OCL/About-OCL/ [Accessed: 2019-05-23].

[45] D. Jordan, J. Evdemon, A. Alves, A. Arkin, S. Askary, C. Barreto, B. Bloch, F. Curbera, M. Ford, Y. Goland, A. Guízar, N. Kartha, S. Commerce, C. K. Liu, R. Khalaf, D. König, M. Marin, V. Mehta, and S. Thatte. Web Services Business Process Execution Language. page 264.

[46] R. R. Mukkamala and IT-Universitetet i København. A formal model for declarative workflows: Dynamic condition response graphs.

[47] V. Diamantopoulou, C. Kalloniatis, S. Gritzalis, and N. Argyropoulos. Supporting the Design of Privacy-Aware Business Processes via Privacy Process Patterns. page 12.

[48] N. Argyropoulos, C. Kalloniatis, H. Mouratidis, and A. Fish. Incorporating privacy patterns into semi-automatic business process derivation. pages 1–12. IEEE. URL http://ieeexplore.ieee.org/document/7549305/ [Accessed: 2018-05-16].

[49] E. Kavakli, C. Kalloniatis, P. Loucopoulos, and S. Gritzalis. Incorporating privacy requirements into the system design process: The PriS conceptual framework. 16:140–158.

[50] N. Notario, E. Ciceri, A. Crespo, E. G. Real, I. Catallo, and S. Vicini. Orchestrating Privacy Enhancing Technologies and Services with BPM Tools: The WITDOM Data Protection Orchestrator. pages 1–7. ACM Press. URL http://dl.acm.org/citation.cfm?doid=3098954.3104057 [Accessed: 2018-03-22].

[51] M. Enamul Kabir, H. Wang, and E. Bertino. A conditional purpose-based access control model with dynamic roles. 38(3):1482–1489. URL http://linkinghub.elsevier.com/retrieve/pii/S0957417410006858 [Accessed: 2018-04-10].

[52] H. Peng, J. Gu, and X. Ye. Dynamic Purpose-Based Access Control. In *2008 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pages 695–700.

[53] N. Y. H. Barringer and N. Zhang. A Purpose-Based Access Control Model. page 8.

[54] J.-W. Byun and N. Li. Purpose based access control for privacy protection in relational database systems. 17(4):603–619. URL http://link.springer.com/10.1007/s00778-006-0023-0 [Accessed: 2018-04-10].

[55] D. Ferraiolo and R. Kuhn. Role-Based Access Controls. In *Proceedings of the 15th National Computer Security Conference*, pages 554–563. URL https://csrc.nist.gov/publications/detail/conference-paper/1992/10/13/role-based-access-controls [Accessed: 2019-05-23].

[56] M. Petković, D. Prandi, and N. Zannone. Purpose Control: Did You Process the Data for the Intended Purpose? In W. Jonker and M. Petković, editors, *Secure Data Management*, volume 6933, pages 145–168. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-642-23556-6_10 [Accessed: 2018-04-10].

[57] A. Lapadula, R. Pugliese, and F. Tiezzi. A Calculus for Orchestration of Web Services. In *ESOP*.

[58] T. Macaulay. Could the Right To Be Forgotten Put The Public Back in Control of Their Data? URL https://www.techworld.com/data/could-right-be-forgotten-put-people-back-in-control-of-their-data-3663849/ [Accessed: 2018-09-17].

[59] Archiveddocs. The STRIDE Threat Model. URL https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v%3dcs.20) [Accessed: 2019-05-23].

Appendix A

GDPR Related Forms

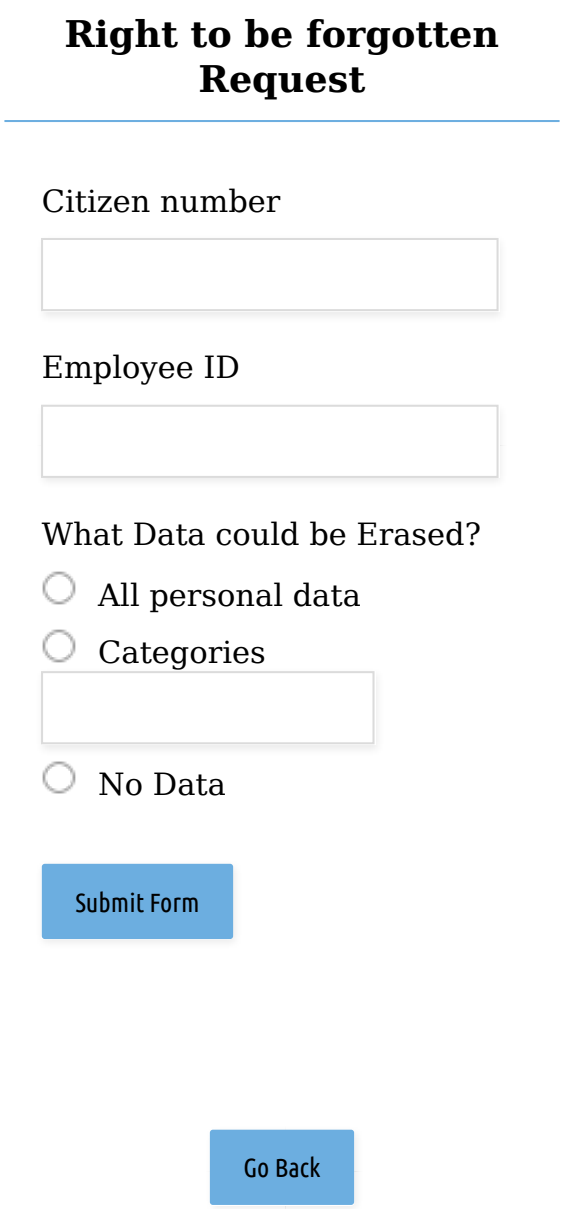

Figure A.1: Form used for exercising the right to be forgotten

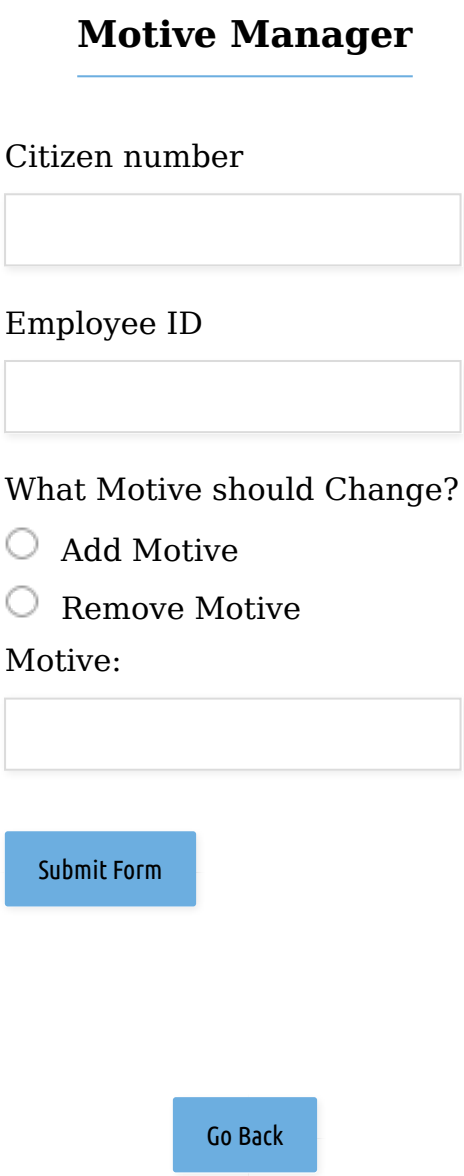

Figure A.2: Form used for updating the list of motives for storing personal data

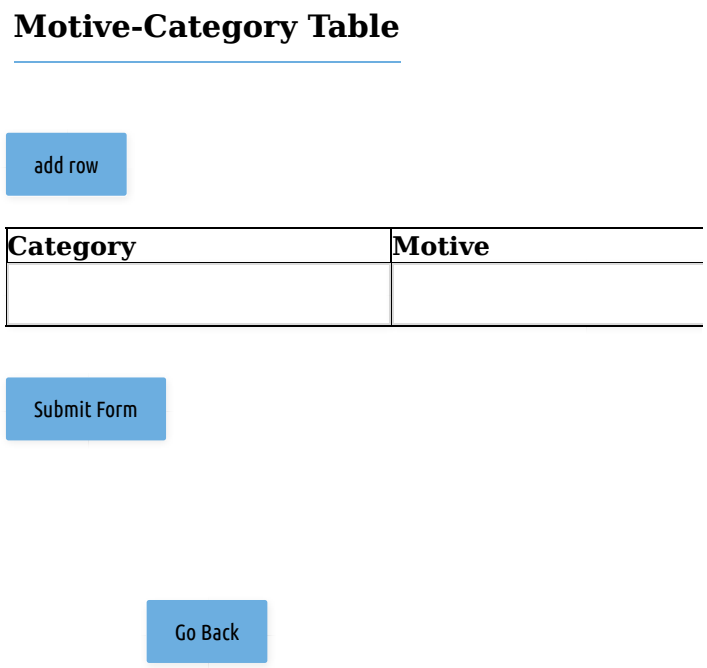

Figure A.3: Form extra used to keep the valid purposes for each data group

Appendix B

Evaluation Results

Parameter collected	Description	#
Processes monitored	Total number of new instances of processes that were monitored	5089
Verified processes	Instances of verified processes that were validated	5035
Failed processes	Instances of failed processes due to incidents	45
Unknown process key events	Number of activities detected for which no running process was found	0
Business rule triggered	Activities that failed due to the business process rule validation failed	43
Activity validators failed	Failed storage minimization activities due to the validator class generated alerts	2
Personal data Informational entities	Instances of the informational entities *Personal data* created	764
Accesses to personal data	Instance of the process Access to Data	4842
Employee data accesses	Number of activities Access to Data for Employee table	1788
Citizen data accesses	Number of activities Access to Data for Citizen table	1647
Other data accesses	Number of activities Access to Data for Other table	1407
Invalid accesses to Employee data	Number of failed Access to Data activities for Employee table	16
Invalid accesses to Citizen data	Number of failed Access to Data activities for Citizen table	21
Invalid accesses to Other data	Number of failed Access to Data activities for Other table	6
Verified accesses to data	Verified Instances of the Access to Data process	4795
Data deletions	Instances of the Remove data process	206
Employee data deleted	Instances of the Remove data process - table '*Employee*'	16
Citizen data deleted	Instances of the Remove data process - table '*Citizen*'	6
Other data deleted	Instances of the Remove data process - table '*Other*'	184
Right to erasure requests	Number of requests to be forgotten detected	15
Verified erasure requests	Number of requests to be forgotten answered	15
Storage minimization validations	Instances of the process Storage minimization	4
Failed storage minimization validation	Instances of the process Storage minimization process in which the activitiy validator failed	2
Added purposes from form	Number activities validated that added a purpose to an informational entity	16
Removed purposes from form	Number activities validated that removed a purpose to an informational entity	6

Table B.1: BP-IDS database values after the evaluation period.