Sensitive Data Discovery and Masking

Rui Miguel Figueiredo dos Santos
IST - Instituto Superior Técnico, Lisboa, Portugal
CINAV - Base Naval do Alfeite, Almada, Portugal
rui.figueiredo.santos@marinha.pt

Abstract— **The technological innovation of the last years, namely related to the automation of the processes of the organizations, has led to the massification of the volume of stored data. Many of the data that the organizations will carry out in their daily activities are, for the most part, data of a personal nature and therefore sensitive. With the entry into force of the general data protection regulation in the European area, it is vital to ensure the protection of data subjects and to ensure that organizations are complying with the rules laid down in the Regulation. The work relates how data discovery can be performed using data mining and machine learning techniques, for the retrieval and extraction of information, either for structured and unstructured data. Extends to the challenges of natural language processing and the identification of open source products that use tools, utilities and libraries (also open source) to understand how they can be applied in the discovery of personal data. The proposed solution is based on the instantiation of a prototype capable of automatically discovering potential sensitive data in the Portuguese language and presenting a model and a prototype that demonstrates the ability to handle the data processing cycle in general.**

Keywords*— *Processing, Privacy and Discovery of personal data. Natural language processing. Structure and unstructured data.

INTRODUCTION

The concern with the privacy of personal data dates back to 1948, with the Universal Declaration of Human Rights [1]. In 1966 an important international provision on privacy emerged, the International Covenant on Civil and Political Rights. The European Union (EU) initiated the legislative process on the protection of the privacy of personal data in 1980 following the OECD Guidelines for the Protection and Privacy of Personal Data. Recently, Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 introduces significant mandatory changes in the operationalization of the principles of protection of personal data.

In the context of the technological revolution, any organization, regardless of the type of business, is connect to the digital world and any system is composed of a wide range of devices (e.g. local servers, mobile devices or IoT). The personal digital data may exist in different formats and supports in the organization. They can be arranged in a *structured* way (e.g. in relational databases, where the fields of the tables have identification and specification of the type of information contained therein); *semi-structured* (*XML* or *RDF* files, in this case with tags allowing the marking of information); and/or *unstructured* (e.g. in *word*, *pdf*, *txt*) thought natural language texts.

The motivation is to identify the main technologies and tools available, which allow the discovery of data of a private nature in organizations. This process involves the identification and location of personal or regulated data, ensuring its proper and safe treatment.

Design Science Research Methodology (DSRM) [2] was used to develop our research, in order to guide the development and evaluation of our proposed method.

The main objective of the work will be to "evaluate and apply open source techniques and tools to discoverer data ensuring the applicability of personal data privacy", from which three specific objectives are derived, as follows: (1) characterization of techniques and tools available for data discovery; (2) instantiate a prototype of tool that integrates capacity in the discovery of data in the Portuguese language; (3) presentation a model and a prototype to demonstrating the capability of the data processing cycle (governing, protecting data and reporting privacy breach incidents).

RESEARCH PROBLEM

Any organization, whether public or private, whether for the purpose of recruitment or the conclusion of a commercial contract, requires the personal data of the person with whom the contract is drawn up, which must be duly safeguarded. In practice, it is often the case that these data are dispersed by the organization, without their access being exclusive, and only to the extent of the needs, to those who need to process them. This naturally hinders the control of data diffusion and ultimately jeopardizes data privacy.

According to the GDPR, any operation developed on the data, namely creation, storage, visualization (consultation), transport, modification, transfer and removal is consider as data processing [3].

Also under the new regulation, the holders have the right to know if a certain organization is treating their personal data in a correct way and to know the purposes of this processing. The holder also has the right to demand that his data be deleted or altered, to request that they not be processed, to oppose their disclosure for marketing purposes, as well as to revoke the consent they have previously granted for certain purposes. The right to the portability of data also gives the holders the privilege of transferring them to another place and having the help/assistance to do so [3].

The GDPR also requires organizations to protect personal data in accordance with their sensitivity. In case of data privacy breach, the data controller should, as soon as possible, notify the competent authorities within 72 hours. In addition, if the breach is likely to pose a high risk to the rights and freedoms of stakeholders, organizations will also have to notify affected individuals immediately [3].

The cycle of processing of personal data begins with the knowledge/discovery and identification of the data and the place where they reside [4]. Substantial management / governance of information, ie how personal data are used and accessed, is then required to be classified as normal and/or sensitive. After classifying the data, it is vital to protect them through security mechanisms, which can be pseudonymized, anonymous or encrypted [3]. The cycle is closed with the notification to the controlling authority or holder and the storage of the necessary records and documents [5].

RELATED WORK

To address the problem presented above, the review of scientific literature has sought data discovery studies in order to identify and understand the latest scientific developments in the field of data discovery in general and personal and/or private data in particular. Although there is a significant volume of studies related to the collection and processing of data, no specific scientific articles have been found specifically related to the discovery of personal data and corresponding techniques used. Most of the literature focuses on the analysis of the data with a view to its transformation into knowledge.

However, a reflection on these techniques has led us to consider that, in a similar way, the techniques to be used for the search and discovery of personal data may be the same as those used to generate knowledge. On the one hand, there are techniques of data mining and machine learning, and on the other, the techniques of retrieval and extraction of information, which have as much of their foundation the previous techniques. In this way, a bibliographical research effort was developed, in order to know the process of discovery of knowledge, with special emphasis on the extraction and retrieval of information, to understand how they are could potentially apply to the discovery of potentially sensitive personal data.

Data discovery

Personal data in digital format can be organized in a structured, semi-structured and/or unstructured form. In this sense, we can divide data discovery (and consequent knowledge) into two approaches: *Knowledge Discovery in Databases (KDD)* and *knowledge Discovery from Text (KDT)*. In the first case, the data are previously organized, whereas in the second approach the data is scattered in text documents or similar.

Structured Data. For Fayyad in 1996 [6] and Mooney in 2005 [7], the processes of data discovery for knowledge generation are composed of several phases, where each phase consists of a set of tasks and each task is solved by means of a technique. The techniques for solving tasks use algorithms, each technique being able to use more than one algorithm. It is an interactive process where all phases play a central role [8]. Each phase is responsible for a slice of the process: (1) Selection; (2) Pre-processing; (3) Transformation; (4) Data mining; (5) Interpretation/Evaluation.

Alternatively, information retrieval (IR) strategies have been adapted in unstructured documents to process, in a freeway, keyword queries on relational databases. There are query models that allow to deal with the problem of multiple word queries, through logical operators AND and OR, allowing the sophisticated exploration of text search in columns [9].

Non-Structured Data. Mooney & Nahm, presented in 2005 [7] a framework for the discovery of data extracted from text *(DiscoTEX)*, whose process is similar to the discovery of structured data. However, according to Miner et al [10], the biggest challenge facing organizations is the identification and treatment of unstructured data. These types of data may be scattered across a wide set of systems, making it difficult to find out. These authors identify and describe seven areas in which text analysis and text mining can help: (1) Search and information retrieval; (2) Document clustering; (3) Document classification; (4) Web mining; (5) Information extraction; (6) Natural language processing; (7) Concept extraction.

The authors advocate that text mining results from the intersection of the seven domains previously mentioned with six main areas of work/development: (1) Data mining; (2) Statistic; (3) Artificial intelligence and Machine learning; (4) Computational linguistics; (5) Libraries and information sciences; and (6) Database.

Information Extraction and Natural Language Processing

The main tasks related to IE [10], which are partly confused with NLP, are Named Entity Recognition (NER), Named Entity Linking (NEL) and Relation Entities (RE).

According Nuno Mamede [11], NLP is extremely difficult to deal with, because some natural languages (NL) are more complicated than others are. E.g., the Portuguese language has much more specific verb tenses than English does, and the conjugation of verbs is more complex. The existence of agreement between words is also a source of problems in many languages. E.g., in Portuguese, nouns and adjectives must agree, which is not the case in English. However, despite the fact that some NLs are particularly complex, they all share the same main problems, which are what make them so difficult to handle from a computational point of view:

- Linguistic variability: the possibility of expressing the same thing in many different ways;
- Ambiguity: the fact that words / expressions / phrases can have several meanings (which leads to more confusion).

Korba et al. [12], in order to automatically discoverer private data in unstructured and/or semi-structured information, concluded that the process necessarily involves IE techniques. Through the NER, the entities, such as names of people, organizations and locations, are located and classified in the texts. In turn, the RE allows identifying semantics of relationship between the entities in the texts. Korba et al. [12] used decision trees, a supervised approach, because it performed well compared to other algorithms.

Tools

The *data defense* tool [13] is embedded in a data discovery program using *Apache OpenNLP* libraries [14]. As key features stand out the ability to identify sensitive personal data, to be an independent platform, able to support links to the Oracle database, MS SQL Server and MySQL and promote it as being useful to assist in the GDPR process.

As pre-requisites, you need JDK 1.8+ and Maven 3+ and the discovery of information related to personal data is based on previously trained binary files that allow you to differentiate different types of data with a certain level of probability using the Maximum Entropy classifier for this purpose. And it uses the *opennlp.tools*, namely *TokenizeME* library to separate words *(tokens)* using the *Maximum Entropy* technique for making decisions and the *opennlp.tools*, trough *TokenizeModel* library that serves to encapsulate the model and promote methods to enable its creation from the binary representation [14].

For the training it should contain several varied sentences, where the words to be trained are recognized as a given entity must be properly labelled with a mark that identifies the type of entity desired and that the training document can be used to train multiple types provided that they are properly marked. It is recommend that the document contain at least 15000 phrases to create a template that has a good level of performance.

Governance/Management

According to Microsoft [4] to be compliance with the GDPR it is vital to have a good model of data governance. After completing the inventory and knowing how personal data are used and accessed, it is important to implement a governance plan that will help to define policies and roles, assign responsibilities for data access, and handle the use of personal data. The definition of a data governance plan, in addition to promoting trust, ensures that the organization effectively respects the legal obligations regarding the privacy of personal data, from its design to transfer or removal.

Protect

Data security, like discovery, is a complex task, since there are several types of risk to identify and consider, from physical intrusions, dishonest employees, accidental loss or computer piracy [4]. Building risk management plans and taking risk mitigation measures, such as password authentication, audit logging and encryption, may be some measures to help ensure compliance with the GDPR. However, it is also recommended, as referred to Pinho [15], the use of data protection anonymization techniques [3]. Data protection measures can be categorized [15] in *randomization* (noise addition, shuffling, differential privacy); *generalization* (k-anonimity, L-diversity) and *pseudonymization* (substitution, encryption, hashing and masking).

Reporting

The GDPR is not limited to the processing of personal data, it establishes new standards of transparency, accountability and record keeping, notably in how organizations should work on documentation that defines internal processes and the use of personal data [4] [3].

Thus, organizations processing personal data should maintain records relating to the purpose of data processing; categories of personal data processed; the identity of third parties with whom the data is shared; (and which) third countries receive personal data and the legal basis for such transfers; technical and organizational security measures; and data retention times applicable to various data sets [4].

Standards and reference models

From the analysis of the GDPR, it can be verified that there is a strong relation with the international norms and that it is recommended that the organizations follow and they implement the recommendations of the normalizations, because their applicability contributes to the confidence of the processes related to the personal data.

One of the main objectives of the ISO is to provide methodologies according to the economic and technical interest, with particular regard to the privacy of personal data and that the GDPR seeks to link through the principles and rights of the data holders, stick to the ISO dedicated to IT security techniques. In order to understand the relationship, it was represent in tree structure:

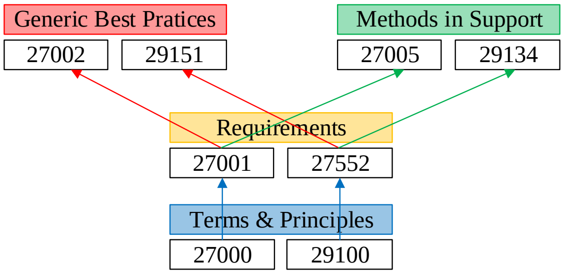

Figure 1 – ISO standards applicable in the GDPR

The COBIT5 model [16] considers that governance and management are two distinct domains because they consider that the activities between them are distinct and serve different purposes in organizational structures: (1) Governance ensures that stakeholder needs, conditions and options are assessed in order to determine agreed and balanced corporate goals; defining direction through prioritization and decision-making; and monitoring performance and compliance with the direction and objectives set. In most organizations, governance is usually the responsibility of the board of directors, under the leadership of the president; (2) The management is responsible for the planning, development, execution and monitoring of activities, in line with the direction defined by the governing body in order to achieve corporate objectives.

The Article 29 Working Party promotes a set of guidelines to ensure compliance with the GDPR. One of the guidelines concerns the DPIA which determines whether the treatment is likely to result in a high risk, stating that different methodologies may be used, giving as an example the international ISO 31000 standard for risk management together with ISO/IEC 29134 for the realization of an DPIA, by recital 90 of the GDPR state several elements that overlap with the defined ones of the risk management. Acknowledging that in the area of risk management, a DPIA is design to manage risks for the rights and freedoms of persons, in particular [17]:

- Establish the context, taking into account the nature, scope, purpose and sources of risk;

- Evaluate the probability or severity of the high risk;
- Respond to risks in order to mitigate risk and ensure data protection.

Research Proposal

The structure of the artefact will be define, with the presentation of its logical architecture, how it will be evaluated, the risk analysis model to be used for data governance and protection measures (mitigation plan).

The aim is to develop the features of the *data defender* tool in order to improve the existing search functionality. For this purpose, the use of dictionaries with names in the Portuguese language and NER learning models for recognition of entities (e.g. proper names, addresses, organizations, etc.) will be used, as well as the construction of regular expression rules identifiers (e.g. tax identification number, telephone numbers and e-mail addresses, etc.). The development of the artefact will be confined to the discovery of sensitive personal data in unstructured and structured environments. It is assumed the existence of access permission, both at the level of the operating system, for the unstructured data, and at the database level, for the structured data. The personal data to be searched has the characteristics defined in the GDPR.

Considering that the isolated discovery of certain attributes does not allow to consider the sensitive data, it is necessary to create classification mechanisms that allow them to relate [12]. However, the primary purpose is to discover personal data, so the device used is intended to be a further warning for the existence of personal data. The classification given in terms of sensitivity and privacy level will result from the relationship of the different types of data existing in a particular context (unstructured document or structured table).

Solution Architecture

The architecture to achieve the dissertation objective was consider a process with four distinct activities, each related to the personal data processing cycle. The detailed description of each activity will be carried out, in the following sections.

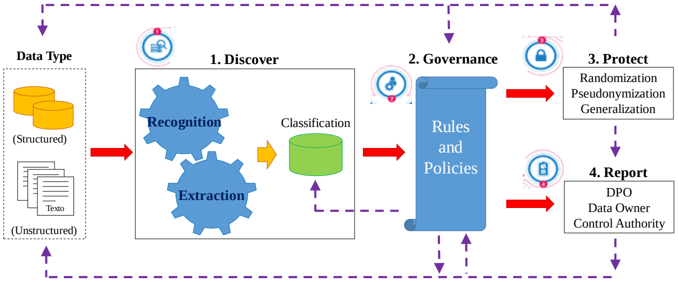

Figure 2: Architecture of propose solution

Discover. The purpose of this activity is to address the specific objective two, obtaining a prototype of the tool with data discovery capability in the Portuguese language, thus improving the functionalities of the data defender tool, in the extraction and retrieval of personal data in Portuguese language, from structured and unstructured data. Similar to Korba et al. [12], dictionaries, standards and regular expressions will be used to allow the application of NLP mechanisms for recognition of terms in the Portuguese language, in order to query the attributes considered sensitive by the GDPR. The classification/importance of the information to be allocated will be in accordance with the criteria defined in the GDPR [3], in particular through Article 5 and Article 9, and NIST SP 800-22 [18] on the confidentiality and privacy of Personally Identifiable Information (PII).

Governance. This activity is a contribution of the specific objective three, to present a model for building a capacity of the data processing cycle. Thus, they propose a set of rules and policies, in accordance with the GDPR and with the best security practices (ISO, NIST and COBIT). The governance process will include more comprehensive actions, because prior to the application of data discovery tools it is vital to know the level of maturity of the organization, i.e. whether there are defined policies for data processing and how they are apply. It is only after the organization's diagnosis has been made that the tools for data discovery should be applied, with the artifact suggesting the most appropriate treatment of the discovered data based on the information discovered.

Protect. The data protection activity is a complement to the previous activity, with the particularity of here one can use techniques of anonymization, pseudonymization and encryption for data protection. Taking the previous example, to the sales area, as this should have access to the name, address and phone, it is recommended that the data accessed are encrypted, and an authentication key is required for access to them. For the statistical area, it is required that the data be pseudonimized.

In view of the complexity and scope of the data security protection domain, the level of security that will be assigned to the prototype is limited to the technological component, based on a set of questions directly related to potential threats will be assigned a mitigation plan encouraging the responsible to carry out the risk analysis exercise responding what is the impact versus probability of a certain event occur.

Reporting. The last activity relates to the third component of the specific objective three of the presented model. The methodology to be developed will focus on aspects related to the production of reports, respecting the changes introduced by the GDPR regarding standards of transparency, accountability and record keeping. Included in these reports are internal to the organization, to the EPD, and external, to the data stakeholders and/or national control authority, depending on the situation.

Proposed Method

Data Classification. Based on the guidelines of the theoretical concepts and recommendations on best practices previously defined in *Standards and reference models*, namely ISO 31000, ISO 27005, ISO 29134 and GDPR standards, the risk matrix for the classification of the risk of personal data.

Considering that the objective of the artifact developed is an alert for the existence of potential sensitive data, the classification matrix proposed as a solution to the problem presents five levels of classification of a certain documents *(unstructured data)* or tables *(structured data).*

Table 1. Level of data risk

Classification			***Description***
1	0,05	Very Low	Data when isolated does not allow identifiable a person.
2	0,25	Low	Combination of more than one of very low personal data (may be identifiable)
3	0,50	Moderate	Data enabling the data subject to be uniquely identified
4	0,75	High	Combination of more than one very low personal data with moderate date.
5	0,95	Very High	Personal data of special categories (whose treatment should be avoided)

Impact Assessment. Associated with the processing of personal data, there are risks that may be of confidentiality, integrity and availability, or even of violation of data subject's rights and/or privacy principles, such as transparency, legitimacy and proportionality. For risk management it is important to identify threats and vulnerabilities, and to assess their impact and likelihood of occurrence [19]. The risk management criteria that will be taken into account for the proposed solution are in accordance with ISO 27005, i.e.: (1) impact; (2) evaluation (probability of occurrence); and (3) acceptance [20].

Based on the classification of risk and according to the data classification, a survey was constructed in order to measure the level of vulnerabilities in information systems, together with the risk density criterion, to calculate risk in the repositories, be structured (database) or in the structure of directories with unstructured data.

For calculating the density of risk of personal data, the following criteria shall be taken into account:

1. Risk classification. According to the criticality of personal data in the repository. This classification will be assigned to the file in case of unstructured data and tables in the case of structured data.
2. Weighting factor. According to the volume of data in the repository. The rationale for this factor is to differentiate documents according to the amount of personal data discovered by assigning a weight based on volume of personal data identified in relation to the total size of the document. That is, it is intended to differentiate documents of identical size but different amount of personal data. In this sense, the weighting assigned is divided into three categories, and for documents containing personal data:

   - Less than 100 is assigned a value of 1;
   - Greater than 100 and less than 1000 is assigned a value of 2; and
   - Greater than 1000 is assign the value 3.

   In order to normalize in a scale between "0" to "1", for the calculation the values are used 0.05, 0.5 and 0.95, respectively.
3. Risk density. Allows you to calculate the risk density of a given document in the repository. The calculation is obtained through the product of risk classification by the weighting factor, on the total of documents in the sample.

$$\sum_{n=1}^{N}\left(\frac{rating\ Data(n) * Weight(n)}{Total\ Sample\ Documents}\right) \qquad (1)$$

Governance/Management. COBIT5 distinguishes governance from management, however, both domains share the need to monitor processes on their accountability [16]. On the other hand, Working Group 29 recommends that responsibility for data processing should be at the highest level of organizations [17]. In this sense, both the board (governance) and the chief executive (management) need to know the overall picture of data privacy that the organization is dealing with.

Thus, there is a need for greater knowledge of existing personal data, of how and where they are stored, who has access, how they are accessed, the type of control mechanisms in existence, the existence of contracts associated with other stakeholders. For this purpose, a data model supporting the discovery of personal data of the artifact was developed.

According to safety standards, we can classify safety levels of three [21]: (1) at a lower or primary level, in physical safety, (2) level of users which includes awareness and training aspects of security and (3) the technological level, through technological knowledge and how it can help in the security, specifically in the protection of personal data, through IT. In this paper, we will only cover technological protection measures.

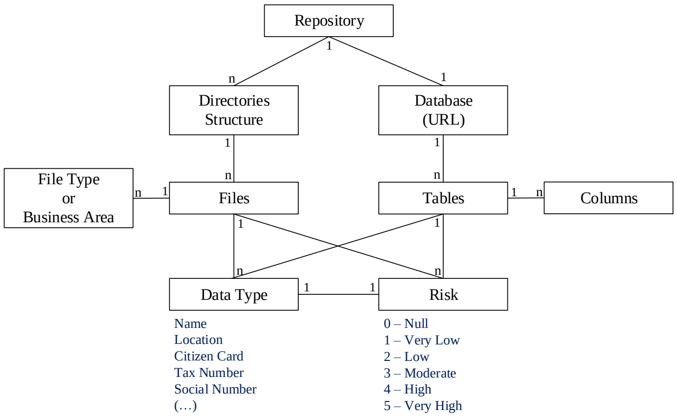

Figure 3: Meta-model to private data governance

In this sense, 30 questions were elaborated on technological security, seeking to address the possible threat / vulnerability. Taking into account the entry into force of Council of Ministers Resolution No. 41/2018 [22], was followed together with the GDPR [3] as a basis for the formalization of the issues because (1) the diplomas meet the standards (reference documents for good practices) and (2) due to the obligation to comply with fines for non-compliance.

The outputs of each question are based on the product of the risk response, with a mitigation plan associated with each question. The sum of all issues will result in the weighted value of the compliance indicator on data privacy.

$$\frac{\sum(Weight\ Responses)}{Total\ Weighted} \qquad (2)$$

Mitigation plan. Following the safety issues mentioned in the previous section, activities considered necessary to meet the minimum level of acceptability of data protection systems were considered. In the implementation of the mitigation plans the objective should be according to the type of data discovered and the business area, according Figure 4, below. On the one hand, threats/vulnerabilities are thrown into issues, and on the other the various action plans, according to the type of threat, with a direct cause/effect relationship with the action to be developed. Also included is the business area that may influence decision making.

Under the proposed mitigation plan, which summarizes the negative responses, data controllers should join with security officers to complete the impact assessment [23]. This is because, non-conforming responses indicate potential threats / vulnerabilities that are not implemented and that can be exploited in some way, contributing in a way to a data privacy violation [3] [18]. Thus, in addition to the given answers, the evaluation should be carried out according to the criteria presented in *Impact Assessment*, considering the impact and probability of occurrence for each question answered negatively. This provides a risk ratio that will serve as an indicator for the data controllers of the criticality level of the measure in relation to the business process.

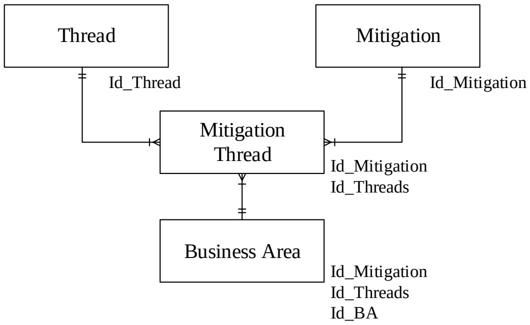

Figure 4: Meta-model of mitigation plan

Finally, the sum of all the risk ratios obtained by the negative response to the total weight of the questions will allow to obtain the risk factor and level of compliance with the GDPR and RCM No. 41/2018.

$$\frac{\sum_{n=1}^{N}(Negative\ Response(n))}{Total\ Answer} \tag{3}$$

IMPLEMENTATION

The instantiation of the prototype consists of a fork of the *data defend*, which explains the factors that motivated the changes made to the referred products and how they were implemented. In addition to the improvements made in the tool (equivalent to first component of the architecture), it is explained how the results were aggregated to serve as input in a new built component, that of governance and protection (second and third component) of the architecture, whose result will promote a summary report (dashboard) of the location of a particular repository (architecture component 4.).

As the main purpose of the developed prototype is to perform the discovery of personal data, the name assigned to the artefact was *PerDa2Disco (Personal Data to Discovery)*.

Creating NLP templates in Portuguese

As the *data defense* tool is develop in the Java language for the construction of new models, it was necessary to use the *OpenNLP* utilities [24] and text annotation scripts to generate the personal data recognition modules for this model. The process for the creation of NL models was carried out in four stages:

1. Identification and selection of texts in Portuguese;
2. Creation of dictionaries with entities to recognize;
3. Note the texts with the appropriate label.
4. Execution of the training template for generating the binary file *(bin)*, for NER.

In addition to the identification of texts written in Portuguese, it was necessary to create annotations in the texts according to the intended entities. To do this, in the first phase a script was created to allow annotations to be made quickly and semi-automatically. The script was developed and used through the *sed* command [25] natively available in *Linux* environments, whose operating principle is based on the reading of the original text written in Portuguese, comparing it with the list of entities intended *(e.g. names)*. Whenever an existing entity in the dictionary matches that of the original text, it automatically enters the annotation around the matching text entity.

After the texts are annotated, the NER training was performed with the *TokenNameFinder* function. Essentially, the program allows you to read the annotated training document and generate the binary files so that the *TokenNameFinder* function allows you to read and execute the generated *bin* file (e.g. pt-ner-name.bin)

New read modes of Data discovery

During the process of creating dictionaries in Portuguese and regular expressions to recognize specific Portuguese data, the thought came to incorporate the ability to analyse documents directly through dictionaries and regular expressions, which is why new libraries were explored and incorporated *(Dictionary)*, dictionary name recognition *(DictionaryNameFinder)*, and regular expression recognition *(RegexNameFinder)*.

In this sense, dictionaries have been changed to *XML* format and several regular expressions have been built, incorporating in the tool the ability to read dictionaries and regular expressions directly, using a process similar to the existing reading for *bin* files, through the function *NameFinder*.

This restructuring of the source code motivated that in the tool, in addition to making the tokenizer with the *NameFinder* library for *Maximum Entropy Model (NERMaxEnt)*, are included new cases to allow the tokenizer for use of dictionaries and regular expressions. The assigned designation for these new discovery modes were *NERDictionary* and *NERRegex*, respectively, and change the native model to *NERMaxEnt*. After the inclusion of these new functionalities, it was verified that the results of the discovery of personal data improved substantially, in addition to presenting superior performance, improvement of data quality was quite significant, as we may have opportunity to verify in the chapter of the evaluation.

Because of linguistic variability (using propositional contractions and articles defined between terms), we explored the possibility of including character-to-character reading in addition to word-reading, creating a new mode of word recognition, called *NERPattern* because based on the comparison of standardized expressions. For the latter mode it was necessary to include new libraries, especially the one of *io.InputStreamReader*.

Based discovery in multiple modes

Although new ways of discovery have been created, the limitation has been that it is necessary to execute the application for the different modes more than once, that is, it is executed with the applicability of dictionaries, sometimes with regular expressions and so on. At the level of operation of the tool and in order to optimize its use, it was developed the possibility to explore in an integrated way the various models simultaneously. The main objective of this functionality is to accelerate the process of discovering personal data, using different models and broadening the spectrum of recognition of entities, as well as simplifying the process of data classification. When choosing more than one model to execute, with just one execution the tool is enabled to evaluate the various parameterized models, returning the best result (according to the weighting given in the search engine) and in case the same location of the document by two or more different.

Dictionaries

Although the dictionaries were initially used to allow the development of new NER models for the Portuguese language, they were eventually retained as a tool search mode. The change that needed to be made was limited to converting the dictionaries to the *XML* format and promoting the changes listed above when creating the new templates.

In total ten different dictionaries have been created that are considered relevant to the topic in question and that in some way the term used in the dictionary has a strong relation with the type of personal data that is to be discovered, in order to univocally identify a certain holder of the data: crimes, localities, marital status, affiliations, gender or sex, education, names, professions, religions and health data.

Regular expressions with validators

In order to allow better flexibility and scalability of the intended rules, the source code has been adapted to accommodate the segregated regular expressions of the source code and to have the ability to read an independent property file. For this functionality we used the *opennlp.tools Regex* utility using the *namefind.RegexNameFinder* [26] library and incorporated validation methods into the source code itself, in case a certain regular expression was detected, namely for the citizen's card (only when the 12 digits), NIF, NISS, credit cards, NIB and Portuguese IBAN (PT50) are used.

The great advantage of including regular expressions with this setting is to allow for modularity and scalability to create regular expressions according to the intended pattern. The only caution to have is that you must abide by the regex rules for the *Java* language type.

In addition to the expressions with validators, expressions were created to detect Portuguese driving license, email, zip code, telephones, mobile phones, and abbreviated dates.

Recognition of compound terms

Another improvement was the introduction of the ability to recognize compound terms, that is, the joining together of more than one consecutive term, with the purpose of increasing the accuracy of the personal data discovered, allowing, for example, in the case of the tool, to discover the Maria and João, if the two terms are followed by each other, it is intended that the result to be returned is only one entity with the name of Maria João and not two entities, one Maria and the other João. are followed, the result should be two distinct entities.

This functionality is only applied in *NERMaxEnt* and *NERDictionary* discovery modes.

Custom query (by pattern model)

Added ability to query through custom standards including compound term searches. In this case, unlike the other models, the query by defaults is not done by means of tokens, but by character to character. The main reason for the development of this functionality was to allow particular consultation, that is, to search for a specific set of data related to a particular person. Essentially, it was thinking of a very particular situation of the GDPR, of a possible claim by a data owner to wonder how and where their personal data is stored.

Search by random sampling

Changed how you begin the process of discovering personal data. Instead of performing the default search on all file types excluding only a few, e.g. *dll* files, has been changed to perform the search by including the files. The reason for this change was to optimize the performance, both localization and analysis, through the parameterization of previously known formats whose content is typically of the text type.

The documents will always be analysed in a random way (it is necessary to set the sample size manually). Regardless of whether the study is conducted by census or by sampling. In the case of the study being done by census (all elements of the population), it would not be necessary to carry out an analysis in a random way, since for the case of the study to be sampled (a small part of all elements of the population), the evaluation is valid and representative of the whole, it is considered that the most appropriate method is by means of a random selection of the documents to avoid tendencies of manipulation of results.

Automation of classification

In order to automate the personal data classification process, a risk grade classification was introduce based to the risk matrix identified in Table 1. As well as logging information, debugging and any errors that may occur (error), the results for further analysis and governance are aggregated into three different files whose extension is of type *csv*, in order to allow different types of analysis:

1. Result. It intends to return all the results found together with some more detailed information, for a possible technical analysis and/or to be able to correlate different types of data to a single proprietor;
2. Summary. It is aggregated according to the filename to allow the evaluation of the risk density of a given repository,

regardless of whether it is structured (database) or unstructured (directory service);

3. Governance. With the objective of being able to help with the EPD of a given organization, to have a more comprehensive view of the dependencies between the various documents and if it is the case of the dependency between repositories, allowing to load this information into another graph visualization tool, simplifying the reading and interpretation of the volume of existing personal data.

To calculate the risk density, all the calculation formulas were added to the code level and a new property file was added in order to allow the parameterization of the document size-weighting factor.

Results presentation

Although it is necessary to use the command line to start the prototype, a graphical *HTML* interface has been developed, using the *Javascript* language, using the W3.CSS framework [27] because it is open source and have a good integrated responsiveness, being very simple and easy to develop and libraries of *jquery.mim.js*, *Chart.min.js* and *Chart.bunble.min.js* for the presentation of the various graphs developed allowing a better visualization of the results found [27].

The results of the discovered data, in the case of unstructured data, were grouped in such a way as to allow a perception of the total volume of existing documents in relation to the sample of analysed data, the density of risk in a particular repository and the time of location and execution. For structured data, the tool indicates the total volume of existing records (total of lines between the various tables) in relation to sample records, risk density and execution time.

EVALUATION

The creation of new NER modules to be executed in the artefact, implies that they are tested in two perspectives: performance and quality. For both cases an evaluation methodology was defined, identifying the metrics to be used and the set of tests to be performed. In a first phase, the tests were carried out in a controlled environment (laboratory) through the creation of scenarios with the existence of properly identified personal data, with the purpose of analysing the results in detail through a reference baseline, and then being able to apply the created models in scenarios close to a production environment (case studies).

The motivation for choosing the case studies was to use the artefact in two organizations with different approaches, but which share the need to process personal data, demonstrating the prototype's functionalities and how it can be advantageous in help in locating the existence of personal data. Thus, two case studies were carried out: the first (1) was in the Portuguese Navy, through search personal data in the *file share* of the IT sector that presents documents distributed between collaborative server and the computers attributed to employees; and the second (2) in data repositories of the Link Consulting company, through *Edoclink system*, because it is a document management application composed by documents and with a lot of information in management metadata, which is why the case study included two distinct components: (a) search in database repositories (structured data), in this case the connectivity used was *Microsoft SQLServer*, and (b) the server where the documents are stored (unstructured data).

Laboratory

The evaluation metrics performed in the laboratory tests consisted in measuring the level of accuracy of the personal data discovered, using metrics commonly used to evaluate the recognition of entities in natural language systems: *Accuracy*, *recall* and *F -measure* and benchmark their performance [28].

For the quality evaluation, 30 documents with different personal information were identified, from different sources and different formats (invoices, curriculum vitae, applications, declarations, among other documents). All documents have been manually reviewed in order to confirm and tag the existing personal data type in order to calculate the relationship between the discovered and actual nominal entities. Together, the documents present 5890 paragraphs and 10178 types of personal data.

Table 2: Quality Metrics Results

Data Type	***Precision***	***Recall***	***F-measure***
Cartão Cidadão	100,00%	91,33%	95,47%
Crime	77,45%	33,76%	47,02%
Email	98,73%	89,66%	93,98%
Estado Civil	71,43%	58,82%	64,52%
Filiação	58,06%	31,03%	40,45%
Habilitação Literária	71,11%	14,35%	23,88%
IBAN	100,00%	93,33%	96,55%
Localidade	47,55%	22,15%	30,22%
Morada	90,32%	43,75%	58,95%
NIB	100,00%	92,00%	95,83%
NIF	95,24%	93,46%	94,34%
NISS	100,00%	90,56%	95,04%
Nome	46,44%	45,23%	45,83%
Profissão	79,59%	53,79%	64,20%
Religião	72,73%	66,67%	69,57%
Saúde	67,14%	34,56%	45,63%
Telefone	96,61%	82,61%	89,06%
Telemóvel	97,18%	90,20%	93,56%

The test consisted of executing the prototype using the various modules created for the different modes of discovery on the 30 previously mentioned documents. Subsequently, the results obtained (Table 2) from the discovery and classification of personal data were compared manually in relation to the previously known documents, verifying the veracity of the same. The process consisted in manually checking all records of the discovered data by marking them according to the metrics mentioned above, allowing them to perform their calculations.

For the same 30 documents, the average performance to search only the attribute “name”, among the different modes of data discovery is according Table 3.

Table 3: Performance Metrics

	Dictionary	***MaxEnt***	***Pattern***
Average per Document (in seconds)	0.40	5.11	20.56

Case Study

Portuguese Navy. According to the interviews and mini-questionnaires made to the main IT actors and responsible for ensuring the application of technological and safety measures in the Portuguese Navy, complemented with interviews, it can be concluded the evaluation in relation to the prototype is generally good. It is not perfect and requires some improvement, especially in some modules for personal data discovery, but the principle is clear and can be an aid to any organization, contributing to the security and privacy of the data, whether to a private or state organization, in general.

The assigned classification of the personal data and the way in which it is attributed its degree was considered a strong point, because it simplifies the process of defining priorities, alerting to the data that could be more problematic. Regarding the results obtained from the analysis made to the IT repositories, the type of personal data discovered is in agreement with the initial one, since it is verified that the types of data discovered correspond to the type of data of the normal functioning of the secretary of the IT sector of the Navy [29].

Link Consulting. It was possible to verify the functioning of the prototype through the discovery of data types in unstructured and structured data, concluding that the prototype presents, good, practical potentialities. Although it needs some improvements, namely in the question of the tuning of false positives, in some models, however given the modularity of the configuration of the modules related to the data types (independent of the tool) it is possible to reduce false positives as long as the models are refined. As strong point of work, they consider (1) the introduction of the concept of Risk Density and (2) how the results are presented, in order to aggregate all relevant information for the execution of protection actions of the data, allowing a prioritization of the most important aspects, facilitating the diagnostic work [30].

Conclusion

Privacy is a fundamental right recognized in the Universal Declaration of Human Rights, and everyone is responsible for ensuring compliance. Technological evolution has led to a massification of data, leading to the emergence of innumerable international norms and recommendations on security, technology and data privacy in recent decades. However, with regard to the applicability of data privacy, there are multiple interpretations at world level, including by different EU Member States. I am convinced that the new GDPR aims at harmonizing the protection of rights and freedoms within the European area by imposing very clear rules on all bodies dealing with personal data.

The aim of this dissertation was to explore IE and NLP techniques to instantiate a prototype capable of detecting, identifying and classifying the different types of personal data, alerting data controllers to becoming aware of the amount of personal data that data different organizational technological repositories (structured and unstructured) may possibly have, thus leading to a reflection on the risks in the data processing process and how they can mitigate the risks of a possible violation of the privacy of personal data.

It was found that the domain of automatic learning, where the IE and NLP techniques used are inserted, is an area in constant innovation and that simply because of the complexity of the natural language domain, its processing is not exactly a simple task, presenting many challenges and difficulties such as language ambiguity and linguistic variability. It is not enough just to make a good application of NER to have a good classification and marking of the entities, it is also necessary to be able to relate them according to the context, only this way you can promote models of superior quality

Future Work

Improve the learning process of the Portuguese models. Although efforts have been made to incorporate the learning process into the prototype, unfortunately it was not possible to implement this functionality. The underlying idea is to label the texts simultaneously to the reading process to analyse if there is personal data where the document analysed would serve to feed *MaxEnt* mode binary files *(bin)* created where successive increments would be made as analyses were performed and validating the classification of the marking of the personal data by means of the results found.

Explore the discovery mode integration by dictionary and by default in order to improve the quality of the accuracy and coverage, especially in relation to duplicate terms, while also seeking to develop standards to optimize the level of performance.

To minimize the number of ambiguities, a field can be create before performing the analysis of the repository to identify itself as to whether to recognize certain terms, possibly according to the nature of the document. That is, if the analysis focuses on a repository that most documents are *curriculum vitae*, we will most likely classify terms such as “Flores”, “Branco”, “Liberdade” or “Céu”, as *name* classification.

Another aspect considered very important, will be to make the tool capable of dealing with preposition contractions (“do”, “de”, “da”, ...), coordinating conjunction (“e”) or definite articles like (“o” or “a”). In this specific domain, the technique of removing *stopwords* can be used, which consists of removing the most frequent words that most often do not reveal information relevant to the construction of the model or, alternatively and for the concrete case of the discovery of personal data, develop a small algorithm to teach in which situations it is important to recognize this kind of words. A possible solution, but not implemented, is to construct a standard module with the standard mode associated with the built algorithm for recognition of compound terms through *Dictionary Mode* reading and *MaxEnt Mode*.

References

[1] United Nations, Universal Declaration of Human Rights, Paris, France: ONU, 1948.

[2] P. Johannesson e E. Perjons, An Introduction to Design Science, Switzerland: Springer International Publishing, 2014.

[3] European Parliament, Regulation (EU) 2016/679 of the European Parliament and of the Council, Brussels: Official Journal of the European Union, 27 april 2016.

[4] Microsoft, Beginning your General Data Protection Regulation (GDPR) Journey, EUA: Microsoft, 2017.

[5] IT Governance Privacy Team 2016, EU GDPR: An Implementation and Compliance Guide, United Kingdom: IT Governance Publishing, 2016.

[6] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth, The KDD Process for Extracting Usefull Knowledge from Volumes of Data, Communications of the ACM, 1996.

[7] R. Mooney e U. Y. Nahm, Text Minind with Information Extration, Bloemfontein, South Africa: Proceedings of the 4th International MIDP Colloquium, 2005.

[8] J. Han e M. Kamber, Data Mining: Concepts and Techniques, San Franscisco, California: Morgan Kaufmann Publishers, 2000.

[9] M. Sayyadian, H. LeKhac, A. Doan e L. Gravano, Efficient Keyword Search Across Heterogeneous Relational Databases, Istanbul, Turkey: ICDE, 2007.

[10] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet e D. Delen, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, USA: Academic Press, 2012.

[11] N. Mamede, “What? Why? Who? Where?,” em *Natural Language Course*, Lisboa, IST, 2017.

[12] L. Korba, Y. Wang, L. Geng, R. Song, G. Yee, A. S. Patrick, S. Buffett, H. Liu e Y. You, Private Data Discovery for Privacy Compliance in Collaborative Environments, Ontario: National Research Council of Canada, 2008.

[13] Armenak, “Sensitive Data Management: Data Discovery and Anonymization toolkit,” GitHub, Inc, 2018. [Online]. Available: https://github.com/armenak/DataDefender. [Accessed on 27 09 2018].

[14] Apache Software Foundation, “Apache OpenNLP Tools 1.9.0 API,” Apache Software Foundation, 2018. [Online]. Available: https://opennlp.apache.org/docs/1.9.0/apidocs/opennlp-tools/index.html. [Accessed on 05 10 2018].

[15] F. A. S. O. Pinho, Anonimização de bases de dados empresariais de acordo com a nova Regulamentação Europeia de Proteção de Dados, Porto: Faculdade de Ciências da Universidade do Porto, 2017.

[16] ISACA, "COBIT 5: A Business Framework for the Governance and Management of Enterprise IT," 2012.

[17] Article 29 Working Party, “WP 248 rev.01 - Guidelines on Data Protection Impact Assessment (DPIA),” 4 10 2017. [Online]. Available: http://ec.europa.eu/justice/data-protection/index_en.htm. [Accessed on 29 04 2017].

[18] NIST SP 800-122, “Guide to Protecting the Confidentiality or Personally Identifiable Information (PII),” NIST - National Institute of Standards and Technology, USA, 2010.

[19] National Privacy Commision, NPC Privacy Toolkit - A Guide for Management & Data Protection Officers, Manila: Department of Information and Communications Tecnhology, 2017.

[20] ISO, “ISO/IEC 27005:2011 - Information technology — Security techniques — Information security risk,” ISO/IEC, Switzerland, 2011.

[21] Presidência do Conselho de Ministros, “Manual de Boas Práticas - Regulamento Geral de Proteção de Dados,” Gabinete Nacional de Segurança, 2018. [Online]. Available: https://www.gns.gov.pt/. [Accessed on 26 09 2018].

[22] Diário da República, 1.ª Série - n.º 62, de 28 de março, Resolução do Conselho de Ministros n.º 41/2018, Lisboa: Presidência do Conselho de Ministos, 2018.

[23] NIST SP 800-53r4, “Security and Privacy Controls for Federal Information Systems and Organizations,” NIST - National Institute of Standards and Technology, USA, 2013.

[24] Apache Software Foundation, “The Apache Software Foundation,” 2018. [Online]. Available: https://www.apache.org/dyn/closer.cgi/opennlp/opennlp-1.8.4/apache-opennlp-1.8.4-bin.tar.gz. [Accessed on 28 09 2018].

[25] E. Pement, “Useful One-Line Scripts for Sed (Unixa Stream Editor),” pemente@northpark.edu, 29 12 2015. [Online]. Available: http://sed.sourceforge.net/sed1line.txt. [Accessed on 05 10 2018].

[26] Oracle, “Package java.util.regex,” Java™ Platform Standard Ed.7, 2018. [Online]. Available: https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html. [Accessed on 20 06 2018].

[27] W3Schools, “w3schools.com - The World's Largest Web Developer Site,” W3Schools, 2018. [Online]. Available: https://www.w3schools.com/js/default.asp. [Accessed on 14 07 2018].

[28] R. Baeza-Yates e B. Ribeiro-Neto, Modern Information Retrieval: the concepts and technology behind search, Second ed., England: Pearson Education Limited, 2011.

[29] Navy IT Stakeholders, Interviewee, *Portuguese Navy Case Study - Evaluate results of prototype.* [Interview]. 20 09 2018.

[30] Edoclink Stakeholders, Interviewee, *Link Consulting Case Study - Evaluate results of prototype.* [Interview]. 09 10 2018.