# Descoberta de dados pessoais, potencialmente sensíveis

**Rui Miguel Figueiredo dos Santos**

Dissertação para obtenção do Grau de Mestre em

## Engenharia Informática e de Computadores

Orientação: Prof. José Manuel da Costa Alves Marques
Eng. Paulo Alexandre Guerreiro Fernandes

**Júri**

Presidente: Prof. Paolo Romano

Orientador: Prof. José Manuel da Costa Alves Marques

Vogal: Prof. Bruno Emanuel Da Graça Martins

**Novembro 2018**

# Agradecimentos

Começo por agradecer com especial ênfase à minha família, em particular à minha ilustre esposa, pela imensa compreensão, paciência e apoio demonstrados, assim como pela privação da minha disponibilidade em determinados momentos.

Ao meu orientador, professor José Manuel da Costa Alves Marques, pela motivação e conhecimentos transmitidos e, finalmente, pelo seu contributo na estruturação, supervisão e revisão da presente dissertação.

Ao professor Anacleto Cortez e Correia (investigador do CINAV – Centro de Investigação Naval), pela sua incansável preocupação, disponibilidade e aconselhamento na fundamentação, procurando conciliar a componente de investigação à componente aplicada.

Ao engenheiro Paulo Fernandes, pelo sentido critico na identificação de potenciais melhorias e no apoio em superar algumas dificuldades sentidas ao longo do percurso de edificação do protótipo, principalmente com a inclusão da agregação por área de negócio e a todos os intervenientes da Link Consulting que de alguma forma contribuíram para tornar realidade o produto implementado.

À Marinha Portuguesa, pela confiança e oportunidade em mim depositada ao ser selecionado para a frequência do mestrado em Engenharia Informática e de Computadores que resultou na promoção da presente dissertação como investigador do CINAV, em particular à Direção de Tecnologias de Informação e Comunicações (DITIC), por intermédio do seu Director Capitão-de-mar-e-guerra Moita Rodrigues, pelo apoio prestado e à disponibilidade dos camaradas da Divisão de Sistemas de Software chefiada pelo Capitão-de-fragata Penim Garcia, pela sua incansável amizade, espírito de camaradagem, entreajuda (partilha de experiências) e imprescindível boa disposição.

Ao Chefe de Divisão e meu tutor, Capitão-de-fragata Penim Garcia um abraço muito sentido de obrigado pela forma como conduziu todo o processo sempre disponível e pronto a colaborar; e claro à camarada e amiga Segundo-tenente Maria Teresa Gama, pela disponibilidade e paciência em me aturar através das explicações ao nível da programação, que sem dúvida foram fundamentais, permitindo-me evoluir e enriquecer os meus conhecimentos.

Ao Chefe da Secretaria dos Serviços Partilhados, Capitão-de-mar-e-guerra Silva Rocha, pelo voto de confiança e disponibilidade em autorizar realizar os testes de avaliação, num ambiente de operação normal, e a todos os seus colaboradores em facultarem o acesso às suas áreas de trabalho.

A todos os entrevistados, pela generosidade em me receberem, despendendo do seu tempo para comigo partilhar informações preciosas e fundamentais para a avaliação do artefacto implementado que muito dignifica e enriquece a dissertação.

Por último, mas não menos importante, a todas as pessoas (camaradas e amigos) no geral, pelo apoio que permitiu a sua concretização possível.

# Resumo

A inovação tecnológica dos últimos anos, designadamente a relacionada com a automatização dos processos das organizações, conduziu à massificação do volume de dados armazenados. Muitos dos dados que as organizações manuseiam na realização das suas atividades diárias são, grande parte, dados de natureza pessoal e por isso de caracter sensível. Com a entrada em vigor do regulamento geral de proteção de dados, no espaço europeu, é vital garantir a proteção dos titulares dos dados e verificar se as organizações estão a cumprir com o preceituado no regulamento.

O trabalho relaciona como a descoberta de dados pode ser realizada recorrendo às técnicas de *data mining* e de *machine learning,* para a recuperação e extração de informação em dados estruturados e não estruturados. Estende-se aos desafios do processamento da língua natural e à identificação de produtos de fonte aberta que utilizem ferramentas, utilitários e bibliotecas (igualmente de fonte aberta) para perceber de que forma é que as mesmas podem ser aplicadas na descoberta de dados pessoais.

A solução proposta é assente na instanciação de um protótipo capaz de descobrir de uma forma automática potenciais dados sensíveis na língua portuguesa através de atributos selecionados e apresentar um modelo e um protótipo que demonstre a capacidade para lidar com o ciclo do tratamento de dados no geral.

Constata-se que o desafio não se restringe apenas à descoberta de dados, é preciso treinar os modelos de processamento de língua natural com vista a alcançar bons resultados. Assim como, é imprescindível aliar a segurança tecnológica à privacidade e utilidade dos dados, envolvendo as áreas de negócio através de aproximações sucessivas ao longo de todo o processo de tratamento de dados pessoais.

**Palavras-Chave:** Tratamento de dados pessoais. Privacidade de dados pessoais. Processamento de língua natural. Descoberta de dados pessoais. Dados estruturados e não estruturados.

# 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.*

*It is observed that the challenge is not restricted to the discovery of data, it is necessary to train the models of natural language processing in order to achieving good results. As well, it is essential to combine technological security with the privacy and usefulness of data, involving the business areas through successive approximations throughout the entire process of processing personal data.*

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

# Tabela de conteúdos

# Índice de figuras

# Índice de tabelas

# Lista de Siglas

| | |
|---|---|
| ADU | Administrador do Domínio de Utilizador |
| AIP | Avaliação de Impacto de Privacidade |
| AIPD | Avaliação de Impacto sobre a Proteção de Dados |
| APO | *Align, Plan and Organize* |
| BAI | *Build, Acquire and Implement* |
| C | Cobertura |
| CC | Cartão do Cidadão |
| CDIACM | Centro de Documentação, Informação e Arquivo Central da Marinha |
| CEO | *Chief Executive Officer* |
| CINAV | Centro de Investigação Naval |
| CNPD | Comissão Nacional de Proteção de Dados |
| COBIT | *Control Objectives for Information and Related Technologies* |
| CP | Código-Postal |
| CRP | Constituição da República Portuguesa |
| D | Dados Descobertos |
| DAGI | Direção de Análise e Gestão da Informação |
| DiscoTEX | *Discovery from Text EXtration* |
| DITIC | Direção de Tecnologias de Informação e Comunicações |
| DSRM | *Design Science Research Methodology* |
| DSS | *Deliver, Service and Support* |
| EDM | *Evaluate, Direct and Monitor* |
| EPD | Encarregado de Proteção de Dados |
| EU | *European Union* |
| FN | Falsos Negativos |
| FP | Falsos Positivos |
| HTML | *HyperText Markup Language* |
| IBAN | *International Bank Account Number* |
| IDE | *Integrated Development Environment* |

| | |
|---|---|
| IE | *Information Extraction* |
| IEC | *International Electrotechnical Commission* |
| INESC ID | Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento |
| IPQ | Instituto Português de Qualidade |
| IR | *Information Retrieval* |
| ISACA | *Information Systems Audit and Control Association* |
| ISMS | *Information Security Management System* |
| ISO | *International Organization for Standardization* |
| IST | Instituto Superior Técnico |
| IT | *Information Technology* |
| JDK | *Java Development Kit* |
| KDD | *Knowledge Discovery in Databases* |
| KDT | *Knowledge Discovery from Text* |
| L2F | Laboratório de sistemas de Língua Falada |
| MaxEnt | *Maximum Entropy* |
| MEA | *Monitor, Evaluate and Assess* |
| MS | *Microsoft* |
| NEL | *Named Entity Linking* |
| NER | *Named Entity Recognition* |
| NIB | Número de Identificação Bancária |
| NIF | Número de Identificação Fiscal |
| NISS | Número de Identificação de Segurança Social |
| NIST | *National Institute for Standardization of Technology* |
| NLP | *Natural Language Processing* |
| OA | Observações Ambíguas |
| OCDE | Organização para a Cooperação e Desenvolvimento Económico |
| OD | Observações Duplicadas |
| OE | Observações Erradas |
| OO | Observações Omissas |

| | |
|---|---|
| P | Precisão |
| PerDa2Disco | *Personal Data to Discovery* |
| PD | Possíveis Dados |
| PII | *Personally Identifiable Information* |
| RDF | *Resource Description Framework* |
| RE | *Relationship Extraction* |
| *Regex* | *Regular Expression* |
| RGPD | Regulamento Geral de Proteção de Dados |
| SI | Sistemas de Informação |
| SNS | Serviço Nacional de Saúde |
| SQL | *Structured Query Language* |
| STI | Superintendência das Tecnologias de Informação |
| TI | Tecnologias de Informação |
| VF | Verdadeiros Negativos |
| VP | Verdadeiros Positivos |
| XML | *eXtensible Markup Language* |

# 1 Introdução

## 1.1 Enquadramento

A preocupação com a privacidade de dados pessoais remonta a 1948, com a Declaração Universal dos Direitos Humanos [1], onde no seu artigo 12.º refere:

> *"Ninguém sofrerá intromissões arbitrárias na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem ataques à sua honra e reputação. Contra tais intromissões ou ataques toda a pessoa tem direito a proteção da lei."*

Em 1966 surge uma importante disposição internacional relativa à privacidade, o Pacto Internacional sobre os Direitos Civis e Políticos, que, no seu artigo 17.º, estabelece o seguinte:

> "*1. Ninguém será objeto de intervenções arbitrárias ou ilegais na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem de atentados ilegais à sua honra e à sua reputação.*
>
> *2. Toda e qualquer pessoa tem direito à proteção da lei contra tais intervenções ou tais atentados.*"

A União Europeia *(EU – European Union)* iniciou o processo legislativo relativo à proteção da privacidade dos dados pessoais, em 1980, no seguimento das Diretrizes da OCDE[1] para a Proteção e Privacidade de Dados Pessoais. No entanto, apenas em 1996 promulgou o primeiro Diploma[2] sobre a proteção de dados, legislação essa que veio a ser transposta para o direito nacional por Lei[3] de 1998 [2]. De referir que a Constituição da República Portuguesa (CRP), através do seu artigo 35.º (Utilização da Informática), salvaguarda o direito à privacidade e as formas adequadas de proteção de dados pessoais [3].

Recentemente, o Regulamento 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção do tratamento de dados das pessoas singulares e livre circulação desses dados, revogou a anterior Diretiva 95/46/CE (RGPD - Regulamento Geral sobre a Proteção de Dados) e a Lei n.º 67/98, de 26 de outubro. A nova legislação do RGPD introduz alterações obrigatórias significativas, ao nível da operacionalização dos princípios de proteção de dados pessoais, cuja entrada em vigor ocorreu no passado dia 25 de maio de 2018, em todos os Estados Membros da EU [4].

1 Organização para a Cooperação e Desenvolvimento Económicos.
2 Diretiva 95/46/EC sobre a Proteção de Dados.
3 Lei n.º 67/98, Proteção de Dados Pessoais.

## 1.2 Motivação

O presente projeto de dissertação de mestrado pretende contribuir para o reforço da proteção de dados, através da identificação das principais tecnologias e ferramentas disponíveis, que permitam a descoberta de dados de natureza privada existente nas organizações. Este processo envolve a identificação e localização de dados pessoais ou regulados, garantindo o seu tratamento adequado e seguro. O tratamento de dados pessoais deve ser uma preocupação permanente dos responsáveis pela gestão da informação das organizações, que para o efeito devem realizar auditorias à informação sensível, incluindo dados confidenciais e informações de identificação pessoal *(PII - Personally Identifiable Information)*. Para além da descoberta dos dados, há que assegurar o seu correto manuseio, garantindo a sua confidencialidade, integridade e disponibilidade [5].

No contexto da atual revolução tecnológica, qualquer organização, independentemente do tipo de negócio, está ligada ao mundo digital e qualquer sistema informático é composto por uma ampla gama de dispositivos (e.g. servidores locais, dispositivos móveis ou IoT[4]) que podem ser disponibilizados por serviços acessíveis em diversas localizações (e.g. *in-house* ou na nuvem). Uma vez que a maior parte da informação das organizações é baseada em suporte digital, cabe ao departamento de Tecnologias de Informação (TI) responsabilidades acrescidas no cumprimento do RGPD, daí ser imperativo que as organizações possuam capacidade de identificação de dados na integra, de forma a garantir práticas adequadas de segurança em conformidade legal [5].

Os dados pessoais digitais, podem existir em diferentes formatos e suportes na organização. Podem estar organizados de forma: estruturada (por exemplo em bases de dados relacionais, em que os campos das tabelas possuem identificação e especificação do tipo da informação nele contido); semiestruturada (por exemplo em ficheiros *XML*[5] ou *RDF*[6], neste caso com etiquetas *(tags)* permitindo a marcação de informações); e/ou não estruturada (por exemplo em ficheiros *word, pdf, txt,* , em textos em língua natural).

## 1.3 Caracterização do problema

O objeto do presente projeto de dissertação de mestrado é a garantia da privacidade dos dados pessoais. Qualquer organização, pública ou privada, seja por motivos de recrutamento ou da celebração de um contrato comercial, necessita de dados pessoais da pessoa com quem estabelece o contrato, dados esses que têm que ser devidamente salvaguardados. Na prática, o que se verifica é que muitas das vezes esses dados encontram-se dispersos pela organização, sem que o seu acesso seja exclusivo, e apenas na medida das necessidades, a quem tem necessidade de os processar. Esta situação, naturalmente dificulta o controlo da difusão dos dados e, em última instância, compromete a privacidade de dados.

---

4 *Internet* das Coisas.
5 *eXtensible Markup Language*.
6 *Resource Description Framework*.

De acordo com o RGPD, considera-se como tratamento de dados qualquer operação desenvolvida sobre os dados, designadamente criação, armazenamento, visualização (consulta), transporte, modificação, transferência e remoção [4].

Também de acordo com o novo regulamento[7], os titulares têm o direito de saber se determinada organização está a tratar os seus dados pessoais de forma correta e conhecer os propósitos desse processamento. O titular tem ainda o direito de exigir que os seus dados sejam excluídos ou alterados, solicitar que não sejam processados, opor-se à sua divulgação para fins de *marketing*, assim como de revogar o consentimento que tenha anteriormente concedido para determinados fins. O direito à portabilidade de dados dá também aos titulares dos mesmos, o privilégio de os transferir para outro lugar e dispor de ajuda/assistência para o fazer [4].

O RGPD exige também que as organizações protejam os dados pessoais de acordo com a sua sensibilidade. No caso de violação à privacidade dos dados, o responsável pelo tratamento de dados, logo que possível, deve notificar as autoridades competentes no período de 72 horas. Além disso, se a violação for suscetível de resultar em risco elevado para os direitos e liberdades dos titulares, as organizações terão também de notificar, sem demora, os indivíduos afetados [4].

O cumprimento do RGPD não pode ser encarado pelas organizações, como uma atividade única, mas sim como um processo contínuo. Para garantir a conformidade com o RGPD, as organizações são encorajadas a promover uma cultura de privacidade para proteger os dados pessoais dos indivíduos que com elas se relacionam [5]. O incumprimento do estipulado pode resultar em penalidades significativas para a organização.

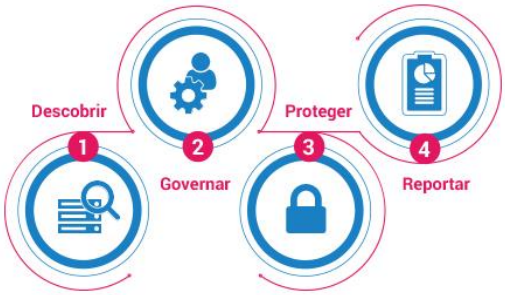



Figura 1.1: Processo do tratamento de dados pessoais

O ciclo do tratamento dos dados pessoais (ver Figura 1.1) é iniciado com o conhecimento/descoberta e identificação dos dados e do local onde residem [6]. Seguidamente é necessária uma gestão/governação rigorosa da informação, ou seja, a forma como os dados pessoais são utilizados e acedidos, classificando-os como normais e/ou sensíveis. Depois de classificar os dados é vital protegê-los por intermédio de mecanismos de segurança, que podem passar pela pseudonimização, anonimização ou cifragem dos mesmos [4]. O ciclo é fechado com a notificação à autoridade de controlo ou titular e o armazenamento dos registos e documentos necessários [5].

7 Regulamento EU 2016/679.

## 1.4 Objetivos do trabalho

O objetivo da presente dissertação de mestrado é o de estabelecer um referencial que contribua para a descoberta de dados pessoais, potencialmente sensíveis, que possam violar a privacidade dos indivíduos, identificando igualmente as formas mais adequadas de tratamento desses dados, de forma a contribuir para o cumprimento das disposições legais estabelecidas no RGPD.

Para o efeito, considera-se que os esforços deverão ser no sentido de analisar os requisitos das disposições legais de uma forma holística, dado que a maioria dos controlos de segurança para a prevenção, deteção, resposta a vulnerabilidades e violações de dados exigidas pelo RGPD são semelhantes a padrões normalizados *(standards)* de proteção de dados já existentes [5].

Deste modo, em vez de endereçar cada norma de forma individual, considera-se que a melhor abordagem para a resolução do problema será identificar as técnicas e tecnologias existentes no mercado, de fonte aberta, analisando e avaliando a forma de como possam ser utilizados para ajudar no processo de conformidade com o RGPD e outros padrões existentes. Adicionalmente, caso seja necessário, pretende-se estender e/ou melhorar a ferramenta selecionada de forma a incorporar técnicas adicionais ou repositórios de dados não suportados.

Em síntese, o objetivo principal do trabalho será o de "avaliar e aplicar técnicas e ferramentas de fonte aberta para descoberta de dados, assegurando a aplicabilidade da privacidade de dados pessoais", do qual derivam três objetivos específicos, a seguir enunciados:

1. Caracterização de técnicas e ferramentas disponíveis à descoberta de dados;
2. Apresentação de um protótipo de ferramenta que integre a referida capacidade na descoberta de dados na língua portuguesa;
3. Apresentação de um modelo e de um protótipo que demonstre a capacidade do ciclo de tratamento de dados (governar, proteger os dados e reportar incidentes de violação da privacidade).

## 1.5 Metodologia

Embora o presente trabalho não seja considerado um projeto puro de investigação cientifica, mas sim um projeto maioritariamente de índole prático procurou-se seguir a abordagem da metodologia *Design Science Research Methodology - DSRM* [7] em virtude de incorporar um conjunto de princípios, práticas e modelos, considerado o mais adequado para o desenvolvimento de projetos na área dos Sistemas de Informação (SI) [8], no caso concreto um projeto de engenharia informática.

Isto porque as soluções para os problemas são encontradas através da criação de novos e inovadores artefactos [9]. Tais artefactos podem incluir construções (símbolos ou vocabulários), modelos (representações), métodos (procedimentos), instanciações (protótipos de sistemas e implementações), inovações sociais, novas propriedades de recursos técnicos, sociais ou informacionais [10] [7]. Por outras palavras, estes artefactos podem ser qualquer objeto projetado com uma solução incorporada

para um problema. Neste sentido, o artefacto proposto é “um método e uma instanciação para a descoberta de dados privados e sua classificação de risco para a privacidade dos mesmos”. A abordagem seguida pelo DSRM inclui seis etapas ou atividades [7]: (1) Identificação do problema e motivação; (2) Definição dos objetivos da solução; (3) Conceção e desenvolvimento; (4) Demonstração; (5) Avaliação e (6) Comunicação.

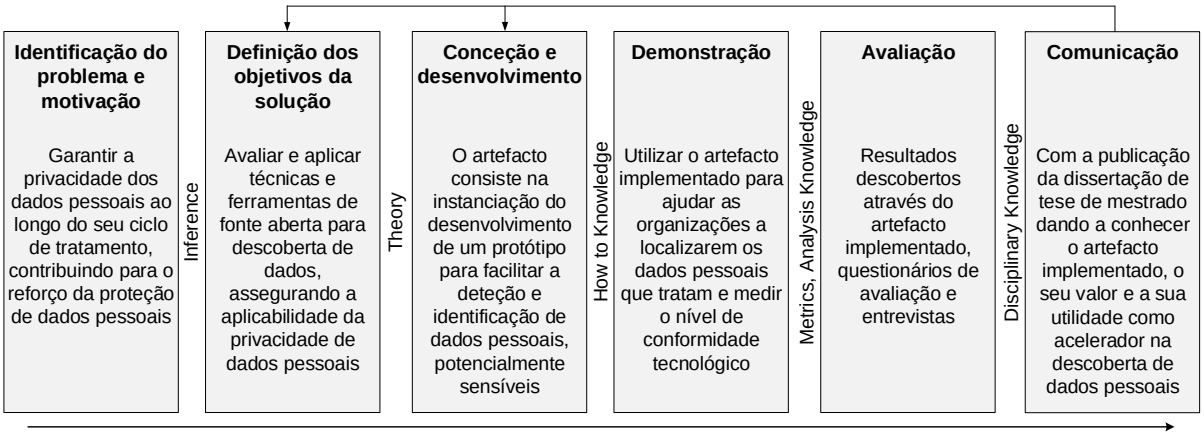



Figura 1.2: Processo DSRM. Adaptado de Peffers [7]

A Figura 1.2, acima, ilustra a instanciação com as seis atividades que orientam o estudo do objeto da dissertação de mestrado, onde se inclui uma pequena descrição de como é que é realizada cada atividade.

## 1.6 Estrutura do documento

A presente dissertação de mestrado, encontra-se estruturada conforme se descreve em seguida.

Inicialmente será apresentado o Trabalho relacionado (Capítulo 2), onde se pretende resumir, com uma breve discussão, as principais técnicas existentes para a descoberta de dados e algumas ferramentas comerciais e de fonte aberta. Será também efetuado um breve enquadramento da importância de governação, principais técnicas de proteção de dados e alterações sobre o reporte de incidentes de violação de incidentes, impostas pelo RGPD.

Seguidamente, no Capítulo 3, será a arquitetura proposta para a solução, bem como as normas e modelos de referência seguidos na análise de risco proposto.

A discussão dos resultados discussão obtidos é efetuada em dois capítulos. O primeiro é focado na implementação do artefacto (Capítulo 4) e o segundo, Capítulo 5, é dedicado à avaliação do artefacto por intermédio dos testes realizados em laboratório (durante a fase de desenvolvimento) e estudos de caso em duas organizações.

A dissertação termina com a síntese das principais conclusões bem como do trabalho proposto.

# 2 Trabalho relacionado

Para endereçar o problema anteriormente apresentado, a revisão de literatura científica procurou estudos relativos à descoberta de dados, a fim de identificar e compreender os últimos desenvolvimentos científicos no domínio da descoberta de dados em geral e de dados pessoais e/ou privados em particular. Embora se verifique um volume significativo de estudos relacionados com a recolha e tratamento de dados, em boa verdade, não foram encontrados artigos científicos relacionados especificamente com a descoberta de dados pessoais e correspondentes técnicas usadas. A maioria da literatura incide o foco da investigação na análise dos dados com vista à sua transformação em conhecimento.

Contudo, uma reflexão sobre as referidas técnicas levou a considerar que, de alguma forma, as técnicas a utilizar para a procura e descoberta de dados pessoais poderão ser as mesmas que as utilizadas para gerar conhecimento. Por um lado, existem as técnicas de *data mining* e *machine learning* e por outro, as técnicas de recuperação e extração de informação, que têm como grande parte do seu fundamento as técnicas anteriores. Deste modo, desenvolveu-se esforço de pesquisa bibliográfica, de forma a conhecer o processo de descoberta de conhecimento e as respetivas técnicas, com especial ênfase na extração e recuperação de informação, processamento de língua natural *(NLP – Natural Language Processing)* e eventuais ferramentas de fonte aberta para perceber de que forma é que as mesmas se podiam aplicar na descoberta de dados pessoais, potencialmente sensíveis.

## 2.1 Descoberta de dados

Como anteriormente referido (no seção 1.2 Motivação), os dados pessoais, no formato digital, podem estar organizados de forma estruturada, semiestruturada e/ou não estruturada. Neste sentido, podemos dividir a descoberta de dados (e consequente conhecimento) em duas abordagens: Descoberta de conhecimento em base de dados (*KDD - Knowledge Discovery in Databases)* e descoberta de conhecimento de texto (*KDT - Knowlwdge Discovery from Text)*. No primeiro caso os dados encontram-se previamente organizados, enquanto que na segunda abordagem os dados estão dispersos em documentos de texto ou similares.

Para *Fayyad* em 1996 [11] e *Mooney* em 2005 [12], os processos da descoberta de dados para a geração de conhecimento são compostos por várias fases, onde cada fase é constituída por um conjunto de tarefas e cada tarefa é resolvida por intermédio de uma técnica. As técnicas para resolução das tarefas recorrem a algoritmos, podendo cada técnica usar mais do que um algoritmo.

### 2.1.1 Dados estruturados.

O processo de *KDD* em bases de dados parece relativamente fácil, pelo facto dos dados já se encontrarem estruturados e organizados. No entanto, segundo *Fayyad et al* [11], neste processo não é

"…trivial a identificação de padrões de dados que sejam válidos e potencialmente úteis.". Figura 2.1, abaixo, representa a visão geral de *Fayyad*, do processo *KDD*.

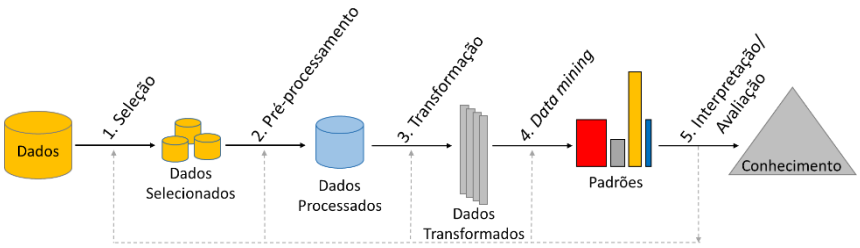



Figura 2.1: Visão geral de KDD. Adaptado de [11]

Como se pode verificar através da Figura 2.1, trata-se de um processo interativo, onde todas as fases desempenham um papel fulcral [13]. Cada fase é responsável por uma fatia do processo, designadamente:

1. Seleção: nesta fase pretende-se a seleção dos dados pertinentes a avaliar. No caso em estudo, os atributos a selecionar contêm potenciais dados pessoais sensíveis;
2. Pré-processamento: eliminação de dados incompletos, isto é, realizar correções e limpeza das bases de dados garantindo a consistência da informação. No presente estudo, esta etapa pretende a redução dimensional da procura dos dados;
3. Transformação: esta etapa é responsável por garantir a persistência dos dados a tratar, preparando-os para o passo seguinte;
4. *Data mining*: pretende procurar padrões através de diversas técnicas existentes nomeadamente, árvores de decisão, regressão e agrupamento;
5. Interpretação/Avaliação: análise do resultado, permitindo ao utilizador a visualização e representação dos dados descobertos.

Não obstante este processo, nas bases de dados, ser relativo à descoberta de dados estruturados (em campos contendo dados atómicos), é também aplicável a textos simples (contidos por exemplo em campos de observações). Embora, os sistemas de gestão de bases de dados forneçam mecanismos de consulta *(query)* de atributos de texto, que incorporam estratégias de classificação de recuperação de informação, essas funcionalidades exigem conhecimento da estrutura da base de dados e que se especifique a(s) coluna(s) às quais uma determinada lista de palavras-chave deve corresponder. Este processo pode ser pesado e inflexível da perspetiva do utilizador. As respostas à consulta por palavras-chave podem precisar da junção de múltiplas relações [14].

Alternativamente, têm sido adaptadas as estratégias de classificação para recuperação de informação *(IR - Information Retrieval)* em documentos não estruturados para processar, de uma forma livre, consultas de palavras-chave sobre bases de dados relacionais. Existem modelos de consultas que permitem lidar com a problemática de consultas de múltiplas palavras, através dos operadores lógicos *AND* e *OR*, permitindo a exploração sofisticada da pesquisa de texto em colunas [15].

*Sayyadian et al*, desenvolveram em 2007 um novo algoritmo *(Kite)* que permite alargar a pesquisa por palavra-chave a múltiplas bases de dados relacionais heterogéneas. O *Kite* combina esquemas das bases de dados e técnicas de descoberta de estruturas para obter junções por chaves (primárias e estrangeiras) das bases de dados envolvidas. Experiências, realizadas com bases de dados em produção, permitiram concluir que o algoritmo é eficiente e produz resultados de elevada qualidade, sem necessidade de recorrer a intervenção humana, ou seja, de forma totalmente automática [15].

### 2.1.2 Dados não estruturados

*Mooney & Nahm*, apresentaram em 2005 [12] uma *framework* para a descoberta de dados com extração a partir de texto (*DiscoTEX*[8]), cujo processo pode ser modelado de acordo com o apresentado na Figura 2.2, apresentaram semelhanças ao da descoberta de dados estruturados.

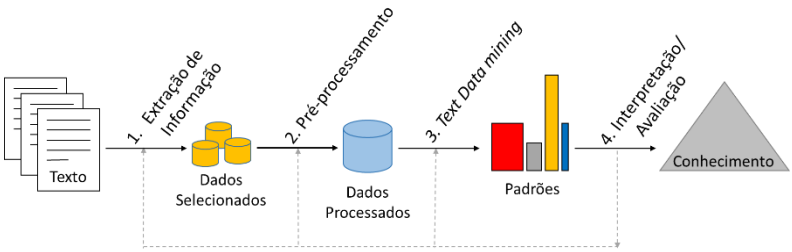



Figura 2.2: Visão geral do processo KDT. Adaptado de [12]

As principais diferenças entre os dois processos estão nas seguintes etapas:

1. Extração de Informação: nesta fase pretende-se selecionar os textos ou palavras chaves de acordo com o domínio do problema, aplicando para o efeito técnicas de extração de informação, tais como o reconhecimento nominal de entidades *(NER - Named Entity Recognition)* e a extração de relações *(RE - Relationship Extration)*;
2. Pré-processamento: a finalidade desta fase do processo é a eliminação de termos não relevantes, tais como pontuações, correções ortográficas e, outros aspetos morfológicos e sintáticos das expressões textuais. Note-se que, no processo KDT, a transformação de dados está incluída na etapa de pré-processamento, cujo resultado deriva na transformação do texto, numa forma estruturada.

Segundo *Miner et al* [16], o maior desafio que as organizações enfrentam consiste na identificação e tratamento dos dados não estruturados. Estes tipos de dados podem estar dispersos por um conjunto alargado de sistemas, o que dificulta a sua descoberta. Estes autores identificam e descrevem sete áreas em que a análise de textos e o *text mining* podem auxiliar, designadamente:

- Pesquisa e recuperação de informação *(Search and information retrieval)*;
- Agrupamento de documentos *(Document clustering)*;

[8] *Discovery from Text EXtration*

- Classificação de documentos *(Document classification)*;
- Mineração Web *(Web mining)*;
- Extração de informação *(Information extraction)*;
- Processamento de linguagem natural *(Natural language processing)*;
- Extração de conceitos *(Concept extraction)*.

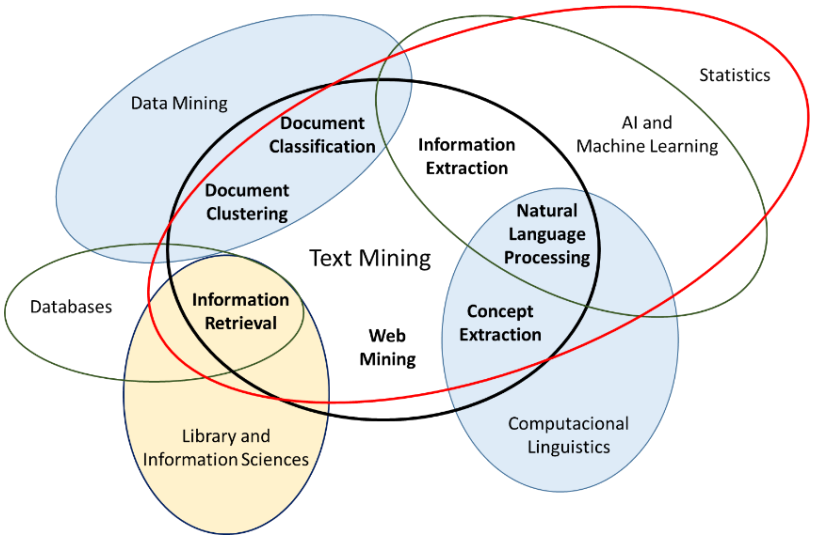



Figura 2.3: Diagrama de interseção dos domínios versus áreas de trabalho. Adaptado de [16]

Conforme se observa na Figura 2.3, o *text mining* resulta da interseção dos sete domínios anteriormente referidos com seis principais áreas de trabalho/desenvolvimento: *data mining*; estatística; inteligência artificial e *machine learning*; linguística computacional; bibliotecas e ciências de informação; e base de dados.

## 2.2 Extração de informação e processamento de língua natural

Nesta seção será dado enfoque à extração de informação *(IE - Information Extraction)*, que em parte se confunde com o *NLP*, descrevendo as principais tarefas subjacentes, as técnicas praticadas e os principais desafios.

As principais tarefas relacionadas com a *IE* são: *NER*; ligação nominal de entidades *(NEL - Named Entity Linking)* e *RE*, que se descrevem seguidamente:

- NER: consiste em identificar e classificar referências a pessoas, organizações, locais e outras entidades num texto. É considerado a componente principal na maioria das aplicações de NLP, nomeadamente na resposta a questões, sumarização e tradução automática [17].
- NEL: é a tarefa de resolver menções a entidades destinadas a serem registadas num base de conhecimento. É muito útil quando se pretende verificar referências a pessoas, lugares e organizações, principalmente em presença de relações ambíguas [18].

- RE: tem como objetivo descobrir ligações semânticas entre entidades. Num texto, equivale geralmente a examinar pares de entidades num documento e determinar (a partir de sugestões no idioma local) a existência de relação entre elas [19].

Outra tarefa importante, quando se analisa dados, nomeadamente para os relacionar (por exemplo, aplicando técnicas de *RE*), consiste em organizá-los de forma estruturada. Para o efeito, é necessário aplicar métodos de classificação de textos, de forma a conseguir a efetiva organização dos dados. Em termos dos algoritmos de classificação de texto, estes estão tradicionalmente divididos em duas técnicas, supervisionadas (preditivas) e não supervisionadas (descritivas) [20]. A diferença entre estas técnicas reside no facto das não supervisionadas não necessitarem de uma pré-categorização dos registos, ou seja, quando não é necessário realizar treino do conjunto de dados a tratar.

Tratam-se de técnicas de *machine learning*, que também podem ser classificadas como técnicas *tradicionais* e *sequenciais*. De entre as técnicas de classificadores sequenciais salientam-se: as *Hidden Markov Models*, *Maximum Entropy Models, Structured Perceptrons*, *Conditional Random Fields* e as *Recurrent or Convolutional Deep Neural Networks.* Para além das técnicas referidas, existem ainda as expressões regulares que consistem numa linguagem de procura de padrões de palavras, através do recurso a regras que permitem associar expressões entre si, formando palavras e/ou números.

Para qualquer das técnicas referidas, o desafio está na formulação dos critérios de procura, que podem ir de simples a múltiplas palavras, listas de interseção e/ou abordagens mais complexas. Segundo Nuno Mamede [21], o NLP é extremamente difícil de lidar, isto porque algumas línguas naturais (NL) são mais complicadas do que outras. Por exemplo, a língua portuguesa possui tempos verbais muito mais específicos do que o inglês, sendo a conjugação de verbos mais complexa. A existência de concordância entre palavras também é uma fonte de problemas em muitos idiomas. Por exemplo, em português, substantivos e adjetivos precisam concordar, o que não é o caso em inglês. No entanto, apesar do facto de que algumas NL são particularmente complexas, todas partilham os mesmos problemas principais, que são os que as tornam tão difíceis de lidar do ponto de vista computacional:

- Variabilidade linguística: a possibilidade de expressar a mesma coisa em muitos modos diferentes;
- Ambiguidade: o facto de que palavras / expressões / frases poderem ter vários significados (o que leva a uma maior confusão).

O Laboratório de sistemas de Língua Falada (L2F) do INESC ID [22] foi criado em janeiro de 2001, integrando a unidade de pesquisa da *Interactive Intelligent Systems*, é constituído por vários grupos de investigadores com um objetivo comum, acrescentar valor agregado ao domínio de computacional de NLP na língua portuguesa. Ao longo destes últimos 18 anos, têm promovido, de entre outras, um vasto número de diretivas para a classificação de diferentes entidades nominais de textos em português, desenvolvendo melhorias constantes nos seus modelos. Como resultado do trabalho desenvolvido, para além do trabalho científico publicado disponibilizam uma versão de demostração do produto *String (A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese)*, a ferramenta compreende uma estrutura modular, capaz de executar todas as tarefas básicas de processamento de

texto, nomeadamente a segmentação de texto *(tokenization)*, marcação de discursos falados, desambiguação morfossintática, análise superficial *(chunking)* e uma análise profunda (extração de dependência) [23].

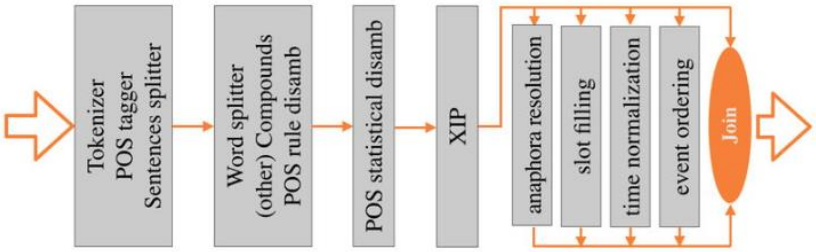



Figura 2.4: Vários Módulos do *String* [23]

*Korba et al.* [24], visando descobrir de forma automática dados privados em informação não estruturada e/ou semiestruturada, concluíram que o processo envolve necessariamente técnicas de *IE*. Através do *NER* localizam-se e classificam-se as entidades, tais como nomes de pessoas, organizações e localizações, nos textos. Por seu turno, o *RE* permite identificar semânticas de relacionamento entre as entidades nos textos. *Korba et al.* [24] utilizaram *decision trees*, uma abordagem supervisionada, por apresentar um bom desempenho comparativamente a outros algoritmos.

A importância de usar simultaneamente estas duas técnicas de extração de informação deve-se ao facto de permitir correlacionar os dados em análise. A descoberta isolada de entidades não permite concluir o carácter sensível dos dados, sendo necessário estabelecer associação com outros dados para se poder aferir do nível de privacidade exigível dos mesmos.

O foco do estudo de *Korba et al.* [24] consistiu na medição do desempenho da extração de informação para três conjuntos diferentes de dados. Os resultados apurados com a comparação dos vários conjuntos de dados indicaram valores ponderados muito positivos de *F-measure*, entre os 72,39% e os 99,41% [24].

# 2.3 Ferramentas

Tendo presente o objetivo da dissertação aliado ao facto de se pretender desenvolver o artefacto com recuso a ferramentas de fonte aberta, o esforço principal centrou-se na identificação de ferramentas que possam servir de ponto de partida para o desenvolvimento do artefacto. Neste sentido, foi identificada a ferramenta ***data defender*** que se baseia em código totalmente aberto e disponível [25], cujo objetivo centra-se na descoberta e anonimização dos dados para permitir conhecer os tipos de dados pessoais em dados estruturados e em dados não estruturados. Os seus responsáveis motivam que qualquer pessoa com interesse na área contribua para melhorar o produto e encorajam que se promovam derivações da ferramenta. Recomendam a sua utilização para realizar auditorias de segurança e consideram que é uma mais valia poder operar o mais próximo possível do ambiente de produção quando se trata de identificar dados sensíveis, contudo alertam que é importante garantir que a sua utilização não comprometa as políticas de privacidade de dados e que estas não sejam violadas [25].

A ferramenta *data defender* [25] está inserida num programa de descoberta de dados com recurso a bibliotecas da *Apache OpenNLP* [26]. Como principais funcionalidades destacam a questão de ser capaz de identificar dados pessoais sensíveis, de ser uma plataforma independente, capaz de suportar ligações a base de dados da *Oracle*, *MS SQL Server* e *MySQL* e promovem-na com sendo útil para ajudar no processo do RGPD.

Como pré-requisitos necessita do *JDK 1.8+* e o *Maven 3+* e a descoberta de informações relacionadas com dados pessoais é baseada em ficheiros binários, previamente treinados que permitem diferenciar diferentes tipos de dados com um determinado nível de probabilidade, utilizando o classificador *Maximum Entropy* para esta finalidade. Os autores disponibilizam um vasto número dos ficheiros binários, mas todos eles foram treinados para a língua inglesa [25]. A página *OpenNLP* [27], disponibiliza igualmente vários modelos pré-treinados mas na língua portuguesa apenas existem modelos de *Tokenizer*, *Sentence Detector* e *POS Tagger*.

Não especificam como é que foram realizados os treinos para a criação dos ficheiros binários, mas de acordo com a informação do *Apache OpenNLP* [26], que disponibilizam alguns modelos e explicam como é que as bibliotecas *OpenNLP* podem ser utilizadas para proceder ao treino dos modelos com recurso à função *TokenNameFinderTrainer* e por intermédio de linha de comandos. Adiantam que o documento de treino deve conter diversas frases variadas, onde as palavras que se pretende treinar sejam reconhecidas como uma determinada entidade têm de estar devidamente etiquetadas com uma marca que identifique o tipo de entidade pretendida e que o documento de treino poderá ser utilizado para treinar múltiplos tipos de dados, desde que estejam devidamente marcadas. É recomendado que o documento contenha, pelo menos, 15 000 frases para criar um modelo que com um bom nível de desempenho. O tipo de etiqueta a utilizar deve estar de acordo com o exemplo ilustrado na Figura 2.5, abaixo.

```
<START:person> Pierre Vinken <END> , 61 years old , will join the board as a nonexecutive director Nov. 29 .
Mr . <START:person> Vinken <END> is chairman of Elsevier N.V. , the Dutch publishing group .
```

Figura 2.5: Exemplo da etiqueta para treino dos modelos [26]

Como referido, a ferramenta *data defender* [25] utiliza a biblioteca *opennlp.tools.tokenize.TokenizeME* para separar palavras *(tokens),* recorrendo à técnica de *Maximum Entropy* para tomar decisões e a biblioteca *opennlp.tools.tonenize.TokenizeModel* que serve para encapsular o modelo e promover os métodos para permitir a sua criação a partir da representação binária [26].

## 2.4 Governação, Proteção e Relatórios

Como se verificou, a tarefa para a descoberta de dados é complexa, carecendo da aplicação de diferentes técnicas, contudo é um passo essencial no processo de tratamento de dados, porque só depois de se conseguir identificar os dados pessoais e saber onde os mesmos residem, é que se está habilitado a determinar o nível de aplicação do RGPD e, eventualmente, a sua conformidade.

Para perceber se o RGPD é aplicável numa determinada organização e quais são as obrigações que lhe são impostas, é importante fazer um inventário dos dados pessoais [6] para ajudar na interpretação dos dados e na identificação dos sistemas onde os dados são recolhidos e armazenados. A realização do inventário permitirá ainda ajudar a entender o propósito dos dados, isto é, como são processados, partilhados e qual o período da sua permanência.

Neste sentido, a Comissão Nacional de Proteção de Dados (CNPD), como entidade administrativa independente [28] que exerce funções de instância nacional de controlo e incorpora o grupo de trabalho n.º 29 para a proteção de dados, que consiste num órgão consultivo europeu independente em matéria de proteção de dados e privacidade, divulgou um documento com as 10 medidas para preparar a aplicação do Regulamento Europeu de Proteção de Dados, que resumem as várias orientações promovidas pelo órgão consultivo europeu, destacando os seguintes tópicos [29]:

1. Informação aos titulares dos dados;
2. Exercício dos direitos dos titulares dos dados;
3. Consentimento dos titulares dos dados;
4. Dados sensíveis;
5. Documentação e registo de atividade de tratamento;
6. Contratos de subcontratação;
7. Encarregado de proteção de dados (EPD);
8. Medidas técnicas, organizativas e de segurança do tratamento;
9. Proteção de dados, desde a conceção e avaliação de impacto;
10. Notificação de violações de segurança.

### 2.4.1 Governação / Gestão

Segundo a *Microsoft* [6] para se estar em conformidade com o RGPD é vital possuir um bom modelo de governação dos dados. É defendido também que, depois de completar o inventário e conhecer como os dados pessoais são utilizados e acedidos, é importante implementar um plano de governação que irá ajudar a definir políticas e papéis, atribuição de responsabilidades para o acesso de dados e manuseamento da utilização dos dados pessoais. A definição de um plano de governação dos dados, para além de promover a confiança, garante que a organização respeite efetivamente as obrigações legais sobre a privacidade dos dados pessoais, desde a sua conceção à transferência ou remoção.

Para a criação de um modelo de governação dos dados pessoais, pretende-se seguir as melhores práticas no que concerne à gestão de segurança da informação, como se poderá observar mais à frente no seção 2.5, Normas e modelos de referência, procurou-se estudar as normas internacionais mais

orientadas para a segurança de informação: as *ISO*[9], as NIST[10] e o COBIT[11] da ISACA[12] com o objetivo de correlacionar com o RGPD para perceber se satisfazem os requisitos do recente normativo europeu.

### 2.4.2 Proteger

Embora nos dias de hoje haja consciencialização nas organizações sobre a importância da segurança de informação [6], o RGPD vem elevar a importância dessa segurança, ao obrigar as organizações a aplicar medidas técnicas e organizacionais apropriadas para a proteção de dados pessoais, estendendo a proteção aos dados perdidos e aos acessos não autorizados [4] [6].

A segurança dos dados, à semelhança da descoberta, é uma tarefa complexa, pois existem vários tipos de risco a identificar e considerar, desde intrusões físicas, empregados desonestos, perda acidental ou pirataria informática [6]. Construir planos de gestão de risco e tomar medidas de redução de risco, como autenticação com recurso a senha, relatórios *(logs)* de auditoria e criptografia, poderão ser algumas medidas para ajudar a garantir a conformidade com o RGPD. No entanto, é também recomendado, como refere Pinho [30], o recurso a técnicas de anonimização na proteção de dados [4]. As medidas de proteção de dados podem ser categorizadas [30] em *randomization (noise addition, shuffling, diferential privacy)*; generalização *(k-anonimity, L-diversity)* e pseudonimização (substituição, encriptação, *hash* e *masking*).

### 2.4.3 Reportar incidentes

Por último, é importante referir que o RGPD, não se limita ao tratamento dos dados pessoais, estabelece novos padrões de transparência, responsabilidade e manutenção de registos, designadamente no modo como as organizações devem atuar relativamente à documentação que define os processos internos e a utilização de dados pessoais [4] [6].

Assim, as organizações que processam dados pessoais deverão manter registos relativos à finalidade de processamento dos dados; as categorias de dados pessoais processados; a identidade de terceiros com quem os dados são partilhados; se (e quais) países terceiros recebem dados pessoais e a base jurídica de tais transferências; as medidas de segurança técnicas e organizacionais; e os tempos de retenção de dados aplicáveis a vários conjuntos de dados [6].

## 2.5 Normas e modelos de referência

Da análise do RGPD, pode-se verificar que existe uma forte relação com as normas internacionais e que é recomendado que as organizações sigam e implementem as recomendações das normalizações, pois a sua aplicabilidade contribui para a confiança dos processos relacionados com os dados pessoais.

9 *International Organization for Standardization*
10 *National Institute for Standardization of Technology*
11 *Control Objectives for Information and Related Technologies*
12 *Information Systems Audit and Control Association*

Para além do processo de conformidade do tratamento de dados ser um processo iterativo e cíclico, é igualmente necessário seguir um conjunto de normas, e que preferencialmente sejam reconhecidas internacionalmente e seguidas por um universo bastante alargado de Estados. A Organização Internacional de Normalização *(ISO)* é uma entidade, criada em 1947, composta por profissionais de todo o Mundo incluindo o Estado português desde 1949, que se faz representar pelo Instituto Português de Qualidade (IPQ) [31]. Um dos principais objetivos das ISO consiste em fornecer metodologias de acordo com o interesse económico e técnico, sendo que no que respeita concretamente a privacidade dos dados pessoais e que o RGPD procura vincular através dos princípios e direitos dos titulares dos dados, cingem-se às ISO dedicadas às técnicas de segurança das TI *(Information technology -- Security techniques)*, destacando-se, de entre outras, as normas abaixo descritas, agrupadas pela sua natureza:

1) Termos e Princípios [32] [33]:
   a) ISO/IEC 27000:2018 *Information security management systems -- Overview and vocabulary*
   b) ISO/IEC 29100:2011 *Privacy framework*
2) Requisitos [34]:
   a) ISO/IEC 27001:2013 *Information security management systems*
   b) ISO/IEC CD 27552 *Extension to ISO/IEC 27001/2 for privacy information management*
3) Boas práticas [35] [36]:
   a) ISO/IEC 27002:2013 *Code of practice for information security controls*
   b) ISO/IEC 29151:2017 *Code of practice for personally identifiable information protection*
4) Métodos de suporte [37] [38]:
   a) ISO/IEC 27005:2018 *Information security risk management*
   b) ISO/IEC 29134:2017 *Guidelines for privacy impact assessment*

De modo a melhor compreender a relação entre as normas, procedeu-se à sua representação em estrutura de árvore como se pode observar na Figura 2.6.

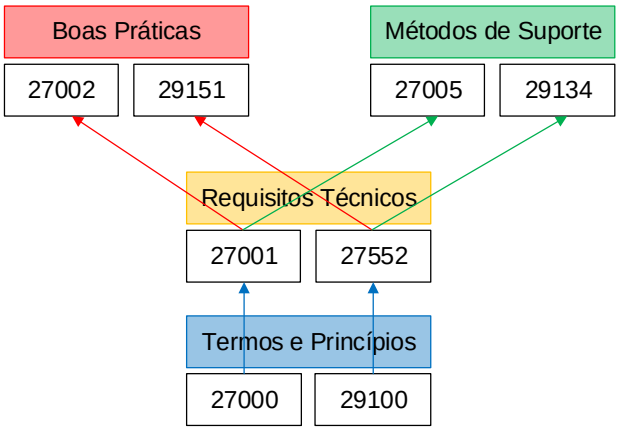



Figura 2.6: Árvore das ISO com aplicabilidade no RGPD

Uma das vantagens da metodologia das ISO é que podem ser aplicadas a qualquer tipo de organização, com vista a analisar o estado de segurança da informação, em virtude de estar focada nos processos e procedimentos, onde a materialização pode divergir ligeiramente de acordo com a

especificidade de cada ambiente tecnológico [34]. Defendem que a sua implementação deve seguir um ciclo contínuo composto por quatro fases: *Plan*, *Do*, *Check* e *Act*, com o objetivo de minimizar os riscos com o cumprimento dos atributos de segurança: confidencialidade, integridade e disponibilidade.

No contexto dos processos de gestão do risco em sistemas de segurança da informação, as atividades a desenvolver em cada fase são [37]:

- *Plan*: Visa estabelecer o contexto, realizar a avaliação do risco, desenvolver o plano de tratamento dos riscos e definir o nível de aceitação do risco;
- *Do*: Implementação do plano de tratamento dos riscos;
- *Check*: Monitorização continua e revisão dos riscos;
- *Act*: Manutenção e melhoramento dos processos de gestão do risco nos sistemas de segurança da informação.


Figura 2.7: Fases de implementação das ISO. Adaptado de [37]

O modelo do COBIT 5 [39] considera que a governação e a gestão são dois domínios distintos em virtude de considerarem que as atividades entre ambos são distintas e satisfazem propósitos diferentes nas estruturas organizacionais, sendo que:

- A governança garante que as necessidades, condições e opções das partes interessadas sejam avaliadas a fim de determinar objetivos corporativos acordados e equilibrados; definindo a direção através de priorizações e tomadas de decisão; e monitorizando o desempenho e a conformidade com a direção e os objetivos estabelecidos. Na maioria das organizações, normalmente a governação é da responsabilidade do conselho de administração, sob a liderança do presidente;
- A gestão é responsável pelo planeamento, desenvolvimento, execução e monitorização das atividades, em consonância com a direção definida pelo órgão de governação a fim de alcançar os objetivos corporativos. Tipicamente, na maioria das organizações, a gestão é da responsabilidade da diretoria executiva sob a liderança do diretor executivo (CEO[13]).

Defendem que as organizações devem implementar os seus processos de acordo com as necessidades de negócio e de forma a que abranjam todas as áreas principais, segregando as atividades dos processos de governação (avaliar, dirigir e monitorizar) e dos processos de gestão

[13] *Chief Executive Officer*

(planear, edificar, executar e monitorizar) [39]. Na Figura 2.8 são ilustradas as principais áreas, assim como as relações existentes entre elas.

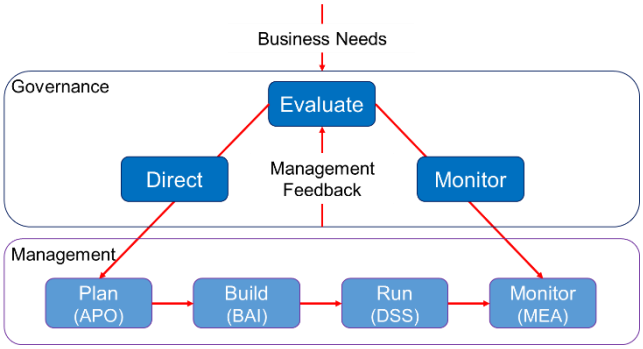



Figura 2.8: Principais áreas do modelo de referência do COBIT 5 [39]

No total identificam 37 processos de governação e de gestão, onde se incluem processos de análise de risco das TI. O domínio da governação contém cinco processos, sendo que em cada processo são definidas práticas de acordo com as áreas anteriormente identificadas (*Evaluate*, *Direct* e *Monitor*). Já o domínio de gestão contém quatro subdomínios, alinhados com as várias áreas responsáveis por planear, edificar, executar e monitorizar, cujos nomes são:

- *Plan:* APO (*Align, Plan and Organise)*;
- *Build:* BAI (*Deliver, Acquire and Implement)*;
- *Run:* DSS (Deliver, Service and Support);
- *Monitor:* MEA (*Monitor, Evaluate and Assess)*.

O motivo da escolha do modelo COBIT 5 surge pelo facto de ser um produto da organização ISACA, reconhecida a nível mundial pelo seu conhecimento, certificação, formação, segurança nos SI, na análise de risco e conformidade das TI. Para além de ser reconhecida a nível mundial, apresenta um modelo único integrado que está alinhado com outros padrões internacionalmente reconhecidos (exemplo as ISO), para além de promover uma arquitetura simples para a orientação da estruturação dos processos de negócio.

Tabela 2.1: Relação entre o RGPD e o COBIT 5

| **RGPD** | **Processos COBIT 5** | |
|---|---|---|
| | **Governação** | **Gestão** |
| Definição dos Riscos e Avaliação de Impacto sobre a proteção dos dados | EDM 02 e 03 | APO 11, 12 e 13 |
| Tratamento dos dados pessoais (Proteger, Processar e Armazenar) | EDM 05 | APO 01, 02, 03 e 10 |
| Consentimento, portabilidade, direito no acesso e a “ser esquecido” | EDM 05 | APO 01, 08, 09 e 10<br>BAI08 |
| Responsabilização do EPD | EDM 01 | APO 07<br>BAI 05 |
| Relato fugas ou violações de dados | | DSS 01, 02, 03, 04, 05 e 06 |
| Assegurar o cumprimento do regulamento | | APO 04, 05 e 06<br>MEA 01, 02 e 03 |

Seguindo o modelo COBIT 5 é possível garantir a conformidade com o RGPD ou melhorar os processos de negócio, por ser viável e relativamente fácil de mapear os vários processos que compõem o COBIT 5 nos princípios e medidas necessárias para garantir a conformidade com o RGPD. A Tabela 2.1, acima, pretende ilustrar a correlação entre o RGPD e os processos do COBIT 5 a implementar.

Por outro lado, o RGPD, no seu artigo 35.º, é claro quanto à aplicabilidade e aos elementos e considerações que devem constar na Avaliação de Impacto sobre a Proteção de Dados (AIPD), assim como na definição do papel e das responsabilidades do EPD. Por este motivo, focando exclusivamente a componente de proteção de dados, mais concretamente dos dados pessoais, pode-se restringir o modelo COBIT 5 ao domínio de governação, verificando-se que está fortemente associado à realização de uma análise e avaliação de risco. Isto significa que o guia de implementação do COBIT 5 poderá ser uma boa ajuda para a definição de atividades a desenvolver, nomeadamente, através dos processos *EDM02 – Ensure Benefits Delivery* e o *EDM03 – Ensure Risk Optimization*.

Em complemento ao referido anteriormente, o grupo de trabalho do artigo 29.º, promove um conjunto de orientações para assegurar a conformidade com o RGPD. Uma das orientações diz respeito à AIPD, que determina se o tratamento é «suscetível de resultar num elevado risco», referindo que podem ser utilizadas diferentes metodologias, dando como exemplo a norma internacional ISO 31000 para a gestão de riscos, juntamente com a ISO 29134 para a realização de um AIPD, em virtude do considerando 90 do RGPD enunciar vários elementos que se sobrepõem aos definidos da gestão de risco. Alertando que em matéria de gestão de riscos, uma AIPD destina-se a “gerir os riscos” para os direitos e as liberdades das pessoas singulares, nomeadamente [40]:

- Estabelecer o contexto, tendo em conta a natureza, o âmbito, as finalidades e as fontes do risco;
- Avaliar a probabilidade ou gravidade do elevado risco;
- Dar resposta aos riscos, de modo a atenuar o risco e assegurar a proteção dos dados.

A interpretação que é feita do motivo pelo qual o grupo referir a ISO 31000 em detrimento da ISO 27005, é porque esta diz respeito às linhas orientadoras gerais para a gestão de risco como um todo, não particularizando ou não se confinando apenas à componente tecnológica. No entanto, como o presente estudo incide na componente tecnológica, a combinação mais adequada será a conjugação da ISO 27005 com a ISO 29134, como métodos de suporte para a criação de listas e/ou questões de conformidade do RGPD, isto porque o que é regulado e solicitado no artigo 35.º, do RGPD, é praticamente análogo a uma Avaliação de Impacto de Privacidade (AIP) da ISO 29134 que por sua vez os processos são similares ao da avaliação de risco da ISO 27005. Neste sentido é possível fazer o mapeamento entre as normas e o RGPD, como se pretende ilustrar através da figura 9.

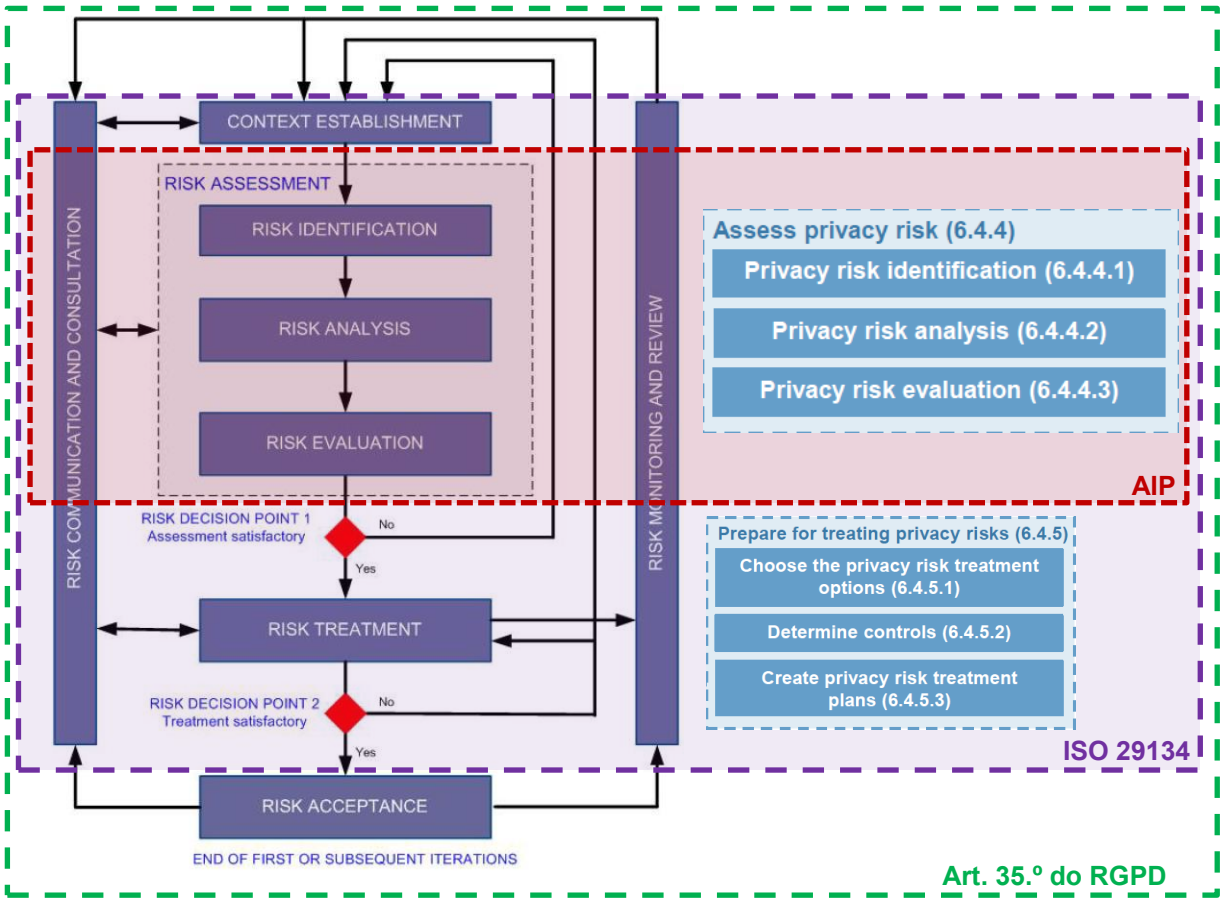



Figura 2.9: Mapeamento ISO 27005 com a ISO 29134 e o RGPD. Adaptado de [37] [38] [4]

## 2.6 Sumário

Embora o artefacto se foque maioritariamente na descoberta de dados, foi considerado pertinente aprofundar a revisão de literatura no domínio do conhecimento, em virtude de se ter verificado a ligação deste a outros domínios, tais como a recuperação e extração de informação estendendo-se ao universo do NLP. Outro motivo que levou ao aprofundamento da revisão da literatura, foi a necessidade de definir a fronteira entre a descoberta de dados estruturados e não estruturados. Finalmente, pretendeu-se, também encontrar semelhanças entre os outros domínios e o processo de descoberta de dados considerados sensíveis.

Foi possível concluir que a descoberta de dados pessoais, potencialmente sensíveis, é uma tarefa complexa e que requer a aplicação de técnicas de diferentes domínios, com muitos desafios relacionados com o NLP, tais como a ambiguidade e variabilidade linguística. Para a aplicação de técnicas de *NER* é preciso de classificar as entidades de textos para poder proceder ao seu treino, podendo ser utilizados dicionários onde deverão constar as entidades pretendidas, como nomes próprios e/ou apelidos para reconhecimento de pessoas, moradas e organizações, assim como, a criação de expressões regulares, para entidades identificadas, com a finalidade de auxiliar no reconhecimento de números (por exemplo o número de identificação fiscal e telefones) e endereços de correio eletrónico ou simplesmente para definir padrões linguísticos de usabilidade. E para a aplicação

de técnicas de *RE*, com forte incidência nos classificadores de texto, com o intuito de conseguir também aferir correspondências semânticas entre as entidades.

Para além das técnicas mencionadas, será igualmente necessário recorrer a ferramentas de *NLP*, tais como a sinalização, segmentação de palavras, interpretadores e, naturalmente, a deteção de língua (*language dection*), com especial enfâse na língua portuguesa, cuja métrica de avaliação será preponderante para perceber o nível de precisão *(precision)* e cobertura *(recall)*. Outro aspeto importante consiste em verificar a capacidade das ferramentas realizarem consultas no domínio de dados estruturados e não estruturados, tendo em conta a particularidade da procura em bases de dados relacionais.

Finalmente, verifica-se um volume significativo de investigação na área NLP com muitas ferramentas criadas e desenvolvidas, com o intuito de automatizar os processos de classificação de entidades, assim como a existência de muitas bibliotecas de fonte aberta nas mais variadas linguagens de programação e muitas demonstrações práticas das potencialidades dos trabalhos científicos desenvolvidos na área. Contudo, são muito poucas as ferramentas totalmente de fonte aberta com a disponibilização de informação rigorosa das suas potencialidades, pelo que apenas foi possível identificar a ferramenta *data defender* que servirá de referência para o modelo proposto.

Com o presente capítulo, fruto de todo o trabalho de investigação relacionado com as técnicas e ferramentas disponíveis para a descoberta de dados, considera-se que se alcançou o objetivo específico um e que estão reunidas as condições para se poder promover uma metodologia de trabalho para o desenvolvimento do artefacto.

# 3 Solução proposta

No presente capítulo, é apresentado o modelo proposto para a recolha e classificação de risco dos dados pessoais. Será definida a estrutura do artefacto, com a apresentação da sua arquitetura lógica, a forma como será avaliado, o modelo de análise de risco a ser utilizado para a governação dos dados e as medidas de proteção (plano de mitigação).

Pretende-se para o efeito desenvolver as funcionalidades da ferramenta *data defender* no sentido de melhorar as funcionalidades de procura existentes. Para o efeito ir-se-á recorrer à utilização de dicionários com nomes na língua portuguesa e modelos de aprendizagem NER para reconhecimento de entidades (e.g. nomes próprios, moradas, organizações, etc.), bem como, à construção de regras de expressões regulares que permitam verificar identificadores (e.g. número de identificação fiscal, números de telefone e endereços de correio eletrónico, etc.). O desenvolvimento da ferramenta será confinado à descoberta de dados pessoais sensíveis em ambientes não estruturados e estruturados. Assume-se a existência de permissão de acessos, quer ao nível do sistema operativo, para os dados não estruturados, quer ao nível da base de dados, para os dados estruturados. Os dados pessoais a procurar possuem as caraterísticas definidas no RGPD.

Tendo em conta que a descoberta isolada de determinados atributos não permite considerar os dados sensíveis, é necessário criar mecanismos de classificação que os permitam relacionar [24]. Contudo, o objetivo principal é a descoberta dados pessoais, pelo que o artefacto usado pretende ser mais um alerta para a existência de dados pessoais. A classificação atribuída, em termos de nível de sensibilidade e privacidade, resultará da relação dos diferentes tipos de dados existentes em determinado contexto (documento não estruturado ou tabela estruturado).

## 3.1 Arquitetura da solução

Descreve-se nesta seção as principais atividades desenvolvidas com vista à obtenção da solução proposta.

A NIST SP 800-22 [41] utiliza o termo PII para descrever qualquer tipo de informação relacionada com dados pessoais, que consiste em “qualquer informação sobre um indivíduo cujo tratamento é efetuado por uma qualquer organização, incluindo (1) qualquer informação que possa ser usada para identificar ou reconhecer univocamente a identidade de um individuo, tais como o nome, número de segurança social, data e local de nascimento, nome dos pais ou registos biométricos; e (2) qualquer informação que esteja relacionada ou que possa ser relacionada a um indivíduo, tais como informação clinica, educacional, financeira e profissão”.

O entendimento no espaço europeu é praticamente idêntico, sendo o RGPD no seu artigo 4.º define dados pessoais com o sendo:

> "… informação relativa a uma pessoa singular identificada ou identificável («titular dos dados»); é considerada identificável uma pessoa singular que possa ser identificada, direta ou indiretamente, em especial por referência a um identificador, como por exemplo um nome, um número de identificação, dados de localização, identificadores por via eletrónica ou a um ou mais elementos específicos da identidade física, fisiológica, genética, mental, económica, cultural ou social dessa pessoa singular."

No sentido da obtenção do objetivo da dissertação, considerou-se um processo com quatro atividades distintas, cada uma delas relacionada com o ciclo do tratamento dos dados pessoais (ver Figura 3.1). A descrição detalhada de cada atividade, será efetuada nas seções seguintes.

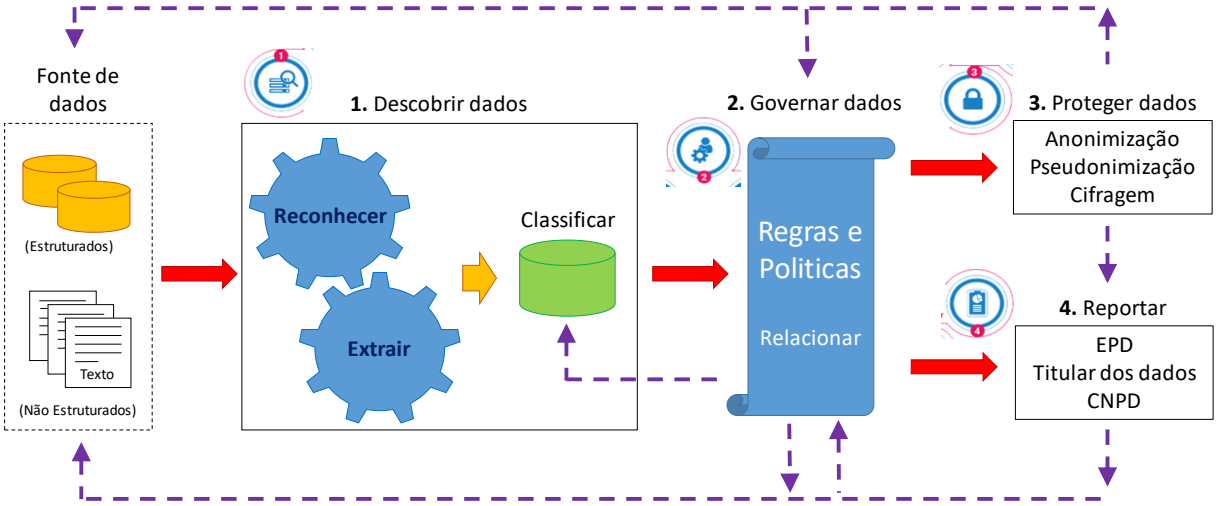



Figura 3.1: Arquitetura da solução proposta

### 3.1.1 Descobrir

Pretende-se com esta atividade endereçar o objetivo específico dois, obtendo um protótipo da ferramenta com capacidade de descoberta de dados na língua portuguesa, melhorando desta forma as funcionalidades da ferramenta *data defender*, na extração e recuperação de dados pessoais em língua portuguesa, a partir de dados estruturados e não estruturados. À semelhança de *Korba et al.* [24], serão utilizados dicionários, padrões e expressões regulares para permitir aplicar mecanismos de NLP para reconhecimento de termos na língua portuguesa, de modo a fazer consultas dos atributos considerados sensíveis pelo RGPD. A classificação da importância/sensibilidade da informação, a atribuir será de acordo com os critérios definidos no RGPD [4], nomeadamente através do artigo 5.º e do artigo 9.º, e pela NIST SP 800-22 [41] sobre a confidencialidade e privacidade de PII, como iremos observar, seguidamente, na seção 3.3, Modelo proposto.

### 3.1.2 Governar

Esta atividade é uma contribuição do objetivo específico três, de apresentar um modelo para a edificação de uma capacidade do ciclo de tratamento de dados. Assim, propõem um conjunto de regras

e políticas, em conformidade com o RGPD e com as melhores práticas de segurança (ISO, NIST e COBIT). O processo de governação irá incluir ações mais abrangentes, pois antes da aplicação das ferramentas para descoberta de dados é vital conhecer o nível de maturidade da organização, ou seja, saber se existem politicas definidas para o tratamento de dados e como é que as mesmas se aplicam. Só depois da realização do diagnóstico da organização é que se deverá aplicar as ferramentas para descoberta de dados, com, em função da informação descoberta, o artefacto a sugerir o tratamento mais adequado aos dados descobertos.

A validação dos dados deverá ter em conta a área de negócio, pois para dados idênticos descobertos poderá haver tratamentos distintos. Por exemplo, a área de vendas poderá ter necessidade de acesso ao nome, morada e telefone para o envio de correspondência, enquanto que a área da estatística não necessita desse detalhe e, portanto, não deve de ter acesso a dados pessoais, potencialmente sensíveis, dos titulares. No entanto a validação terá de ser, forçosamente, efetuada recorrendo a questionários e entrevistas com os responsáveis do tratamento de dados juntamente com os responsáveis pelo SI.

### 3.1.3 Proteger

A atividade de proteção de dados é um complemento da atividade anterior, com a particularidade de aqui se pode usar técnicas de anonimização, pseudonimização e cifragem para a proteção de dados. Pegando no exemplo anterior, para a área de vendas, como esta deverá ter acesso ao nome, morada e telefone, é recomendado que os dados acedidos estejam cifrados, e seja necessária uma chave de autenticação para o acesso aos mesmos. Já para a área de estatística requer-se que os dados estejam pseudonimizados.

Face a complexidade e abrangência do domínio da proteção de segurança dos dados, o nível de segurança que será atribuído ao protótipo cinge-se apenas à componente tecnológica, como se irá poder verificar mais à frente na seção 3.3, onde com base num conjunto de questões diretamente relacionadas com potenciais ameaças será atribuída um plano de mitigação incentivando o responsável a fazer o exercício de análise de risco respondendo qual é o impacto versus probabilidade de um determinado evento ocorrer.

### 3.1.4 Reportar

A última atividade prende-se com a terceira componente do objetivo específico três do modelo apresentado. A metodologia a criar dará enfoque aos aspetos relacionados com a produção de relatórios, respeitando as alterações introduzidas pelo RGPD no que concerne aos padrões de transparência, responsabilidade e manutenção de registos. Incluem-se nesses relatórios, os internos à organização, para o EPD, e os externos, para os titulares dos dados e/ou autoridade de controlo, dependendo da situação.

Como se pode verificar, as últimas três etapas da arquitetura proposta contribuem para o mesmo objetivo, pelo que inicialmente, será apresentado as principais normas e modelos de referência que de

alguma forma podem e estão associadas ao RGPD para permitir propor um modelo mais específico para o tratamento de dados pessoais.

## 3.2 Avaliação

Face ao problema identificado, para dar resposta aos objetivos específicos dois e três (ver página 4), foram definidas como atividades principais do projeto: (1) desenvolver um software que possa contribuir para a descoberta de dados pessoais, potencialmente sensíveis, suscetíveis de violar a privacidade de indivíduos, e (2) identificar uma metodologia para o tratamento de dados, de forma a garantir o cumprimento das disposições legais estabelecidas pelo RGPD. A avaliação dos resultados alcançados, terá de ser capaz de aferir a eficácia dos modelos desenvolvidos para a descoberta de dados pessoais e permitir conhecer o nível de aceitação da implementação dos requisitos funcionais do artefacto. Assim, para aferir a eficácia do tratamento de dados, os métodos de avaliação foram segregados em duas componentes, uma por intermédio de testes em laboratório e outra através de estudos de caso em duas organizações distintas (Marinha e Link Consulting).

### 3.2.1 Laboratório

Os testes em laboratório consistem maioritariamente em permitir medir o nível de exatidão dos dados pessoais descobertos, através de métricas de avaliação padronizadas. Isto é, serão edificados um conjunto de documentos com vários tipos de dados pessoais previamente conhecidos para serem usados como linha referencial para permitir, conhecer a precisão e cobertura dos vários módulos desenvolvidos para os diferentes modos de leitura. Para além das métricas de avaliação serão realizados testes de desempenho temporal de execução do artefacto de acordo com os diferentes modelos edificados.

De acordo com os resultados obtidos na fase de laboratório será construída a base de referência para os estudos de caso a analisar.

### 3.2.2 Estudos de caso

A ideia dos estudos de caso prende-se com o facto de poder testar a ferramenta o mais próximo do ambiente de produção, como sugerido pelos autores da ferramenta *data defender* [25], onde se desconhece o tipo de dados existentes nos repositórios de dados e os mecanismos de segurança implementados.

A análise comparativa será obtida mediante o referencial dos testes em laboratório. Por um lado, pretende-se aferir a diferença dos tempos de execução, em função da localização dos documentos, bem com da análise dos documentos de acordo com os diferentes modelos. Por outro lado, pretende-se conhecer como se comporta o artefacto no decorrer do seu funcionamento e os resultados face ao expectável pelas partes interessadas (responsáveis pela definição das politicas e regras) das organizações.

Neste sentido, serão usados questionários e entrevistas para avaliar o nível de satisfação dos utilizadores e responsáveis pelo tratamento de dados, de duas organizações (Marinha e Link Consulting). Com este instrumento, para além de se pretender medir o nível de satisfação com a ferramenta desenvolvida, pretende-se também conhecer o valor da informação resultante da sua utilização (quer ao nível de funcionalidades quer ao nível de vulnerabilidades), de forma a recolher informação sobre o tipo de preocupações de organizações em diferentes áreas de negócio.

## 3.3 Modelo proposto

### 3.3.1 Classificação dos dados

Com base nas orientações dos conceitos teóricos e recomendações sobre as melhores práticas anteriormente definidas na seção 2.5 Normas e modelos de referência (ver página 15), nomeadamente as normas ISO 31000, ISO 27005, ISO 29134 e RGPD, foi desenvolvida a seguinte matriz de risco para a classificação do risco dos dados pessoais.

Tabela 3.1: Nível de risco dos dados

| Classificação | | | Descrição |
|---|---|---|---|
| 1 | 0,05 | Muito Baixa | Dados que quando isolados não permitem a identificação do titular dos dados |
| 2 | 0,25 | Baixa | Combinação de mais do que um dado pessoal com classificação 0,05 - baixa (o titular dos dados poderá ser identificável) |
| 3 | 0,50 | Moderada | Dados que permitem a identificação unívoca do titular dos dados |
| 4 | 0,75 | Alta | Combinação de mais do que um dado pessoal de classificação 0,05 com o nível 0,50 (permitem a obtenção do perfil do titular dos dados) |
| 5 | 0,95 | Elevada | Dados pessoais de categorias especiais (cujo tratamento deve ser evitado…) |

Tendo em conta que o objetivo do artefacto desenvolvido consiste num alerta da existência de potenciais dados sensíveis, a matriz de classificação proposta como solução para o problema, apresenta cinco níveis de classificação de determinado documento (dado não estruturado) ou tabela (dados estruturados). Conforme referido na seção 3.1.1, Descobrir, a classificação do tipo de dados descobertos está de acordo com a interpretação do RGPD [4], a criticidade da identificação unívoca de um individuo é definida no NIST SP800-22 [41], o que permite construir uma matriz de análise do risco [38] dos dados pessoais. Os níveis de classificação são atribuídos numa escala percentual de modo a normalizar os cálculos.

Assim, considera-se como nível muito baixo (0,05) os dados pessoais genéricos que isoladamente muito dificilmente poder-se-á afirmar referentes a dados pessoais de um titular de dados específico, tais como: nome, localidade, estado civil, género, profissão, habilitações académicas, entre outras. À medida que se agrega outros dados do mesmo documento (não estruturado) ou tabela (estruturado) a

probabilidade de conseguir identificar um individuo aumenta. No caso de se conseguir relacionar o nome com a morada, a profissão ou outro dado genérico a probabilidade de conseguir identificar univocamente um individuo vai aumentando mantendo-se, mesmo assim, no nível baixo (0,25). Já quando estamos perante dados unívocos a classificação por si só será considerada moderada (0,50), pois estamos perante um dado único que permite identificar um titular de dados. Por sua vez se conseguirmos relacionar dados genéricos com um dado unívoco, a probabilidade de poder ter um perfil caraterizador de determinado titular é alta (0,95). Finalmente, quando estamos perante dados que segundo o artigo 9.º, do RGPD, (e.g. origem étnica, saúde, convicções políticas, etc.) não devem ser tratados salvo em situações muito concretas, a ferramenta atribuiu automaticamente uma classificação de nível elevado. Tal acontece mesmo que não esteja relacionado diretamente com um individuo, servindo de uma chamada de atenção para o fato de que se detetaram dados que não devem ser tratados e há necessidade de investigação para perceber o contexto do termo identificado.

### 3.3.2 Avaliação de impacto

A norma portuguesa ISO 31000, define o risco como sendo um “efeito (um desvio, positivo ou negativo, relativamente ao esperado) da incerteza na consecução dos objetivos”, referindo-se igualmente que “o risco é frequentemente expresso como a combinação das consequências de um dado evento e a respetiva probabilidade de ocorrência [42]. Contudo, alerta que para a conceção e a implementação dos planos de gestão do risco é preciso ter em conta as diversas necessidades de uma determinada área específica.

Sendo o risco também avaliado pelo potencial que um incidente poder resultar num dano ou perigo para o(s) titular(es) dos dados [38] foi também analisado o guia de implementação da ISO 29134 para a avaliação de impacto de privacidade (PIA). Aí se estabelece que a análise dos riscos sobre a privacidade dos dados, deve centrar-se em três pontos essenciais: (1) quebra de confidencialidade, através de acessos não autorizados à PII; (2) quebra da integridade, com modificação não autorizada da PII; e finalmente (3) ausência de disponibilidade dos dados, quando se está perante uma perda, roubo ou remoção não autorizada.

A ISO 27005, por seu lado, refere que a abordagem da gestão do risco depende do âmbito e dos objetivos pretendidos e que diferentes abordagens podem ser aplicadas, propondo uma abordagem em que sejam incorporados critérios de gestão de risco: (1) de impacto; (2) de avaliação (probabilidade de ocorrência); e (3) de aceitação [37].

Resumindo, associados ao tratamento de dados pessoais existem riscos que podem ser de confidencialidade, integridade e disponibilidade ou, inclusive, ou mesmo de violação dos direitos dos titulares dos dados e/ou princípios de privacidade, como a transparência, legitimidade e proporcionalidade. Para a gestão do risco é importante identificar as ameaças e vulnerabilidades, e avaliar o seu impacto e probabilidade de ocorrência [43]. Estabelecem-se seguidamente os três critérios da gestão de risco que serão tidos em conta para a solução proposta:

– A nível do **impacto** [43], foram considerados cinco níveis, numa escala percentual para a normalização dos cálculos, sendo que os valores atribuídos foram de acordo com o guia PMBOK [44]. O nível inferior (0,05) diz respeito a um impacto que se possa considerar como desprezável com consequências de fácil superação. Um impacto máximo ou extremo tem um valor de 0,80, registando-se quando as consequências são consideradas catastróficas e impossíveis de ultrapassar em virtude de não ser possível reverter a situação.

Tabela 3.2: Impacto do evento. Adaptado de [43] [44]

| Nível | Impacto | Descrição do impacto | Consequência do impacto |
|---|---|---|---|
| 0,05 | Desprezável | Não serão afetados ou encontrarão apenas alguns inconvenientes | Fácil de superar |
| 0,1 | Limitado | Poderão encontrar inconvenientes significativos | Ultrapassáveis (poderá apresentar algumas dificuldades) |
| 0,2 | Razoável | Poderão encontrar consequências significativas | Difíceis de ultrapassar |
| 0,4 | Significativo | Poderão encontrar consequências significativas irreversíveis | Eventualmente, não se conseguirá ultrapassar |
| 0,8 | Máximo | Irão encontrar consequências significativas irreversíveis | Poderá não ultrapassar |

– Como critério de avaliação da **probabilidade de ocorrência** [43], foram igualmente considerados cinco níveis, cuja escala foi também baseada no guia PMBOK [44]. No nível inferior (0,1) onde não se identifica ou reconhece a possibilidade de ocorrer uma ameaça. À medida que a escala progride a probabilidade de uma ameaça ocorrer vai-se tornando possível mais plausível, até um nível máximo (0,90), relativo a situações muito prováveis.

Tabela 3.3: Probabilidade de ocorrência. Adaptado de [43] [44]

| Nível | Probabilidade | Descrição da Probabilidade |
|---|---|---|
| 0,1 | Quase nula | Possibilidade remota de ocorrer. |
| 0,3 | Pouco provável | Não é expectável, mas existe uma pequena possibilidade de vir a ocorrer. |
| 0,5 | Possível | Ocorrência casual.<br>Em determinadas situações poderá eventualmente ocorrer. |
| 0,7 | Provável | Ocorrência frequente.<br>Existe uma forte possibilidade que venha a ocorrer. |
| 0,9 | Elevada | Muito frequente. A possibilidade de ocorrencia é quase certa.<br>É esperado que ocorra na maioria das circunstâncias. |

– Por fim os **critérios de aceitação** seguem as diretivas do PMBOK para resposta ao risco [44] aliada às práticas de privacidade dos dados [43]. A matriz de probabilidade versus impacto, da Figura 3.2 estabelece as respostas em quatro escalões. A Tabela 3.4 sintetiza os diferentes níveis de resposta a dar.

| Probabilidade (P) | | | | | | |
|---|---|---|---|---|---|---|
| | 0,9 | 0,045 | 0,090 | 0,180 | 0,360 | 0,720 |
| | 0,7 | 0,035 | 0,070 | 0,140 | 0,280 | 0,560 |
| | 0,5 | 0,025 | 0,050 | 0,100 | 0,200 | 0,400 |
| | 0,3 | 0,015 | 0,030 | 0,060 | 0,120 | 0,240 |
| | 0,1 | 0,005 | 0,010 | 0,020 | 0,040 | 0,080 |
| | | 0,05 | 0,1 | 0,2 | 0,4 | 0,8 |
| | | | | Impacto (I) | | |

Figura 3.2: Matriz de Probabilidade versus Impacto. Adaptado de [44]

Tabela 3.4: Resposta ao risco. Adaptado de [44]

| Nível | Tipo | Resposta ao Risco |
|---|---|---|
| 0,005- 0,015 | Insignificante | Aceitar; Transferir |
| 0,020 - 0,080 | Limitado | Aceitar; Reduzir; Transferir; Partilhar |
| 0,100 - 0,240 | Significativo | Eliminar; Reduzir; Transferir; Partilhar |
| 0,280 - 0,720 | Elevado | Eliminar; Reduzir |

Mediante a classificação de risco elencada e de acordo com a classificação de dados, foi construído um inquérito de modo a permitir aferir o nível de vulnerabilidades existentes nos sistemas de informação para juntamente com o critério de densidade de risco[14] se poder calcular risco existente nos repositórios, seja estruturados (base de dados) ou na estrutura de diretórios com dados não-estruturados.

Neste sentido, para o cálculo da densidade de risco dos dados pessoais serão considerados os seguintes critérios:

1) Classificação do risco de acordo com a criticidade de dados pessoais existentes no repositório.
   Esta classificação será atribuída ao nível do ficheiro para o caso da análise ser sobre dados não-estruturados e/ou ao nível das tabelas no caso dos dados estruturados.
2) Fator de ponderação de acordo com o volume de dados existentes no repositório.
   O racional deste fator consiste em diferenciar os documentos de acordo com o volume de dados pessoais descobertos atribuindo um peso de acordo com o volume de dados pessoais identificados em relação ao volume total do documento. Isto é, pretende-se diferenciar documentos de dimensão idêntica, mas quantidade de dados pessoais diferentes.
   Por exemplo, dois documentos com 1000 palavras no total, mas o primeiro documento apresenta 100 possíveis dados pessoais e o segundo apresenta 600 possíveis dados pessoais, a probabilidade do risco de ambos os documentos deverá ser distinta, pois, o documento com mais

[14] Entende-se por **Densidade de Risco** como sendo uma métrica de avaliação para a existência de dados pessoais num determinado repositório.

dados pessoais apresentará uma maior probabilidade de permitir relacionar dados pessoais e eventualmente identificar univocamente um individuo. Para o exemplo, teríamos um rácio de 0,05 e 0,50, respetivamente.

Neste sentido, a ponderação atribuída está repartida em três categorias, sendo que para documentos que contenham dados pessoais:

- Inferiores a 100 é atribuído o valor 1;
- Superiores a 100 e inferiores a 1000 é atribuído o valor 2; e
- Superiores a 1000 é atribuído o valor 3.

De modo a normalizar numa escala percentual entre 0 a 100, para o cálculo são utilizados os valores 0,05; 0,5 e 0,95, respetivamente, conforme se pode observar na matriz representada através da Tabela 3.5, abaixo.

Tabela 3.5: Matriz do fator de ponderação

| Dados pessoais descobertos | Ponderação |
|---|---|
| ≤ 100 | 0,05 |
| > 100 e ≤ 1 000 | 0,50 |
| > 1 000 | 0,95 |

3) Densidade do Risco.

Permite calcular a densidade de risco de determinado documento *(n)* existente no repositório. O cálculo é obtido através do produto da classificação de risco pelo fator de ponderação, sobre o total de documentos[15] na amostra.

$$\sum_{n=1}^{N}\left(\frac{ClassificaçãoDados(n)*Ponderação(n)}{TotalDocumentosAmostra}\right) \quad (1)$$

De modo a simplificar o processo de cálculo do risco, foi considerado definir como requisito a inclusão no artefacto dos cálculos e devolução do valor da densidade de risco de determinado repositório. Para permitir a leitura e interpretação dos resultados descobertos pela ferramenta, foi criada uma interface gráfica em HTML e *javascript* para fornecer ao EPD e demais intervenientes no processo de tratamento de dados os dados pessoais descobertos.

Caso se pretenda analisar mais do que um sistema de diretórios de dados não estruturados, o resultado da densidade de risco engloba todas as estruturas analisadas. Em alternativa, caso o responsável de tratamento de dados tenha acesso a vários servidores de dados partilhados poderá efetuar análises da densidade de risco dos dados pessoais dos vários repositórios.

Adicionalmente à densidade de risco, é importante conhecer as vulnerabilidades a nível tecnológico para se conhecer o desvio existente e, consequentemente, estabelecer um plano de mitigação para

[15] O termo documento significa: nos dados não estruturados equivale a ficheiros e nos dados estruturados equivale a tabelas.

garantir a segurança da informação. Na seção seguinte este aspeto será endereçado, bem como racional do respetivo cálculo.

### 3.3.3 Governação/Gestão

Como referido em 2.5 Normas e modelos de referência (página 15), o COBIT 5 distingue a governação da gestão, contudo, ambos os domínios partilham a necessidade de monitorizar os processos sobre a sua responsabilidade [39]. Por outro lado, o grupo de trabalho 29.º recomenda que a responsabilidade pelo tratamento de dados deva ser do mais alto nível das organizações [40]. Nesse sentido, quer o conselho de administração (governação) quer o diretor executivo (gestão) têm necessidade de conhecer o panorama geral da privacidade dos dados que a organização trata.

Assim, há necessidade de ter maior conhecimento dos dados pessoais existentes, de como e onde estão armazenados, quem tem acesso, como são acedidos, o tipo de mecanismos de controlo existentes, a existência de contractos associados a outros intervenientes. Para o efeito foi desenvolvido um modelo de dados de suporte à descoberta de dados pessoais da ferramenta PerDa2Disco (ver Figura 3.3).

Cada repositório poderá ser composto por estruturas de diretórios (dados não estruturados) e/ou por base de dados (dados estruturados). Por sua vez uma **estrutura de diretórios** é constituída pela raiz *(root)* e por várias pastas e subpastas de diretórios, cuja representação assume a forma de árvore. A sua complexidade e definição do nível de permissões de acesso depende de organização para organização, motivo pelo qual no meta-modelo apresentado é afirmado uma relação de *um* repositório para *n* estruturas de diretórios. Independentemente disso a estrutura de diretórios poderá ter mais do que um ficheiro (documentos) e estes podem ter diferentes tipos de formato e/ou de negócio; já no caso das **bases de dados** a razão perante um repositório é de *um* para *um*, ou seja, como iremos observar no estudo de caso da Link Consulting (seção 5.2.2, ver página 64), tipicamente as organizações associam uma base de dados a apenas um repositório (servidor) onde incluem uma estrutura de diretórios para armazenamento de documentos. A base de dados poderá ser composta por várias tabelas e estas poderão ter várias colunas com um determinado tipo. Independentemente de se estar perante uma estrutura de diretórios ou de uma base de dados, ambos poderão ter vários tipos de dados e associado ao tipo de dados será apresentada uma classificação de risco.

Em termos práticos e de modo a ser possível uniformizar os diferentes tipos de dados (estruturados e não estruturados), a estrutura de diretórios será representada ao mesmo nível da base de dados; os ficheiros ao mesmo nível das tabelas e, por último, o formato dos ficheiros estão ao mesmo nível das colunas.

Para uma melhor perceção da relação dos tipos de dados pessoais foi considerada a visualização da mesma através de grafo, em virtude de permitir uma melhor compreensão das dependências e das relações existentes. A ferramenta selecionada para esse efeito foi aplicação EAPY[16] que consiste numa

[16] https://admin.eapy.eu/

ferramenta de governação de dados, desenvolvida pela empresa Link Consulting, capaz de gerir de uma forma rápida repositórios de governação de dados, escalável e permite uma visualização interativa baseada em grafos [45].

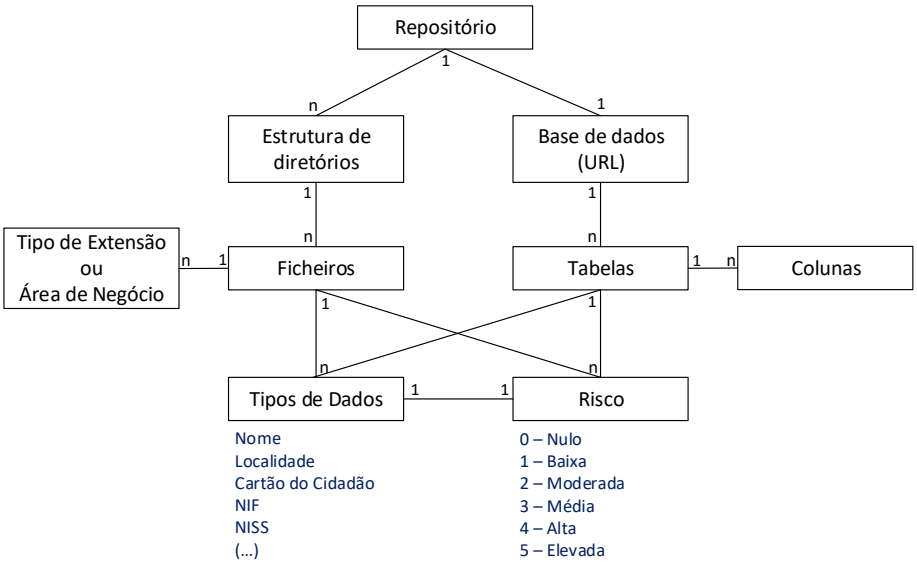



Figura 3.3: Modelo de meta-dados para recolha de dados para a governação de dados pessoais

Adicionalmente ao modelo usado para se ter uma visualização gráfica da relação dos dados pessoais, foi realizada uma comparação entre o COBIT5 e o RGPD de forma a mapear a relação existente entre ambos e permitir a elaboração de questões sobre a avaliação de conformidade.

No Apêndice A são apresentadas 80 possíveis questões comuns entre o COBIT5 e o RGPD [46] [4]. Isto significa que na maioria dos artigos do RGPD é possível formalizar uma questão que está diretamente relacionada com uma ou mais atividades de um determinado processo do COBIT5, seja ele um processo de governação ou de gestão. Os processos que mais contribuíram para o mapeamento efetuado são os apresentados na Tabela 2.1: Relação entre o RGPD e o COBIT 5 (ver página 18) [39].

De acordo com as normas de segurança, podemos classificar os níveis de segurança de três [47]: (1) num nível mais baixo ou primário, na segurança física, (2) nível dos utilizadores onde se inclui a consciencialização e formação dos aspetos de segurança e (3) o nível tecnológico, através do conhecimento tecnológico e como o mesmo poderá auxiliar na segurança, concretamente na proteção dos dados pessoais, através das TI. Neste trabalho apenas iremos abordar as medidas de proteção ao nível tecnológico.

Neste sentido, no Apêndice B é apresentado um conjunto de questões sobre segurança tecnológica procurando equacionar a possível ameaça/vulnerabilidade. Tendo em conta a entrada em vigor da Resolução de Conselho de Ministros n.º 41/2018 [48], seguiu-se o modelo apresentado no anexo A do mesmo diploma juntamente com o RGPD [4], como base para a formalização das questões porque (1) os diplomas atendem às normas standards (documentos de referência para boas práticas) e (2) devido à obrigatoriedade de cumprimento com coimas por incumprimento.

As questões elaboradas, têm resposta binária (sim ou não), ou uma escala entre 1 a 5. O resultado de cada questão é baseado no produto da resposta pelo risco, com um plano de mitigação associado a cada questão. O somatório de todas as questões resultará no valor ponderado do indicador de conformidade sobre a privacidade de dados.

$$RácioConformidade = \frac{\sum(PesoRespostas\ Questões)}{TotalPesoQuestões} \quad (2)$$

### 3.3.4 Medidas de Proteção

Na sequência das questões sobre segurança tecnológica referidas na seção anterior, foram consideradas atividades necessárias para satisfazer o nível mínimo de aceitabilidade dos sistemas de proteção de dados. Na implementação dos planos de mitigação o objetivo deverá estar de acordo o tipo de dados descobertos e da área de negócio, conforme a Figura 3.4, abaixo. Por um lado, as ameaças/vulnerabilidades são vertidas em questões, e por outro os vários planos de ação, de acordo com o tipo de ameaça/vulnerabilidade, com uma relação direta de causa/efeito com a ação a desenvolver. É também incluída a área do negócio que poderá influenciar a tomada de decisão.

No Apêndice C são apresentadas as várias atividades/ações a desenvolver na área de segurança tecnológica [34] com especial ênfase para a privacidade de dados [36] [41] na vertente de dados pessoais, ao abrigo do RGPD [4].

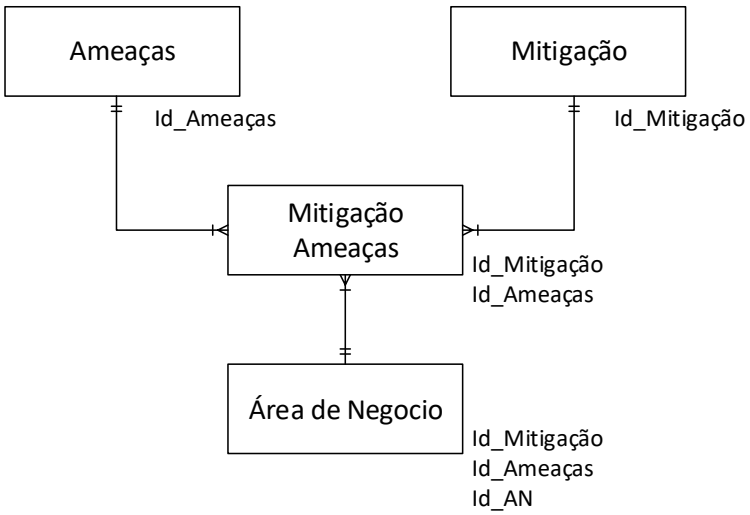



Figura 3.4: Modelo de dados do plano de mitigação

Mediante o plano de mitigação proposto, que resume as respostas negativas, os responsáveis pelo tratamento de dados devem se juntar com os responsáveis pelo domínio da segurança para completarem a avaliação de impacto [49]. Isto porque, as respostas não conformes indicam potenciais ameaças/vulnerabilidade que não estão implementadas e que de alguma forma poderão ser exploradas, contribuindo de certa forma para uma violação de privacidade de dados [4] [41]. O somatório de todos os rácios de risco obtido pelas questões *(n)* cuja resposta foi negativa em relação

ao peso total das questões, permitirá obter o fator de risco e nível de conformidade com o RGPD e a RCM n.º 41/2018.

$$FatorRisco = \frac{\sum_{n=1}^{N}(PesoRespostaNegativa(n))}{PesoTotalQuestões} \quad (3)$$

Finalmente, em complemento às respostas dadas, quando é apresentado o quadro resumo final, o responsável é convidado a realizar uma avaliação de impacto tecnológico de acordo com os critérios apresentados na seção 3.3.2, considerando o impacto e a probabilidade de ocorrência (ver Tabela 3.2 e Tabela 3.3, página 29) para cada questão respondida negativamente (aspetos não conformes). Obtém-se assim um rácio de risco que servirá como indicador para os responsáveis pelo tratamento de dados pessoais do nível de criticidade da medida em relação ao processo de negócio, devendo a resposta à medida estar de acordo com a Tabela 3.4 (ver página 30). Esta avaliação é recomenda que sejam unidos esforços dos responsáveis da área de negócio juntamente com os responsáveis das TI.

## 3.4 Sumário

Para a apresentação do modelo proposto, foi efetuado uma analise detalhada ao RGPD e da forma como o regulamento segue recomendações internacionais em matéria de segurança sobre a privacidade de dados e como é que a segurança deve e pode ser utilizada para garantir a sua conformidade.

Concluiu-se que, de facto, existe uma relação forte entre as normas internacionais e o regulamento, verificando-se que as organizações que atualmente apresentem uma cultura de segurança e sigam as recomendações das normalizações internacionais para implementação dos processos terão, certamente, um elevado nível de conformidade com a privacidade de dados. Outra metodologia muito completa e vocacionada para a implementação de processos de acordo com as necessidades de negócio é o COBIT 5, constatando-se que uma das suas grandes vantagens consiste no facto de conseguir agregar vários tipos de recomendações internacionais transformando-as em atividades e/ou linhas de ação com um nível de granularidade adequados às áreas de negócio.

Foi possível verificar e confirmar que o RGPD assenta nas melhores práticas sugeridas das normas internacionais destinadas quer à segurança (através das ISO 27001, ISO 27002 e ISO 27005) quer à segurança de gestão da privacidade de dados (através das ISO 27552, ISO 29151 e ISO 29134). Arriscaria a afirmar que um dos objetivos principais do RGPD consiste em forçar a aplicabilidade das recomendações existentes sobre a matéria de segurança e privacidade de dados, através da criação de um diploma legal e com a previsão de aplicação de coimas para quem não o aplicar.

Face ao exposto, foi possível construir um modelo que respeita o ciclo do tratamento de dados que se considera que poderá ser útil como acelerador para reconhecer e identificar a existência de dados pessoais e detetar eventuais desvios de comportamentos quanto à existência de dados pessoais dispersos, alertando os principais responsáveis para os perigos e riscos de uma possível violação de dados pessoais. A metodologia proposta servirá de base para permitir desenvolver o artefacto

contribuindo parcialmente para aos objetivos específicos dois e três que serão complementados e finalizados com a implementação e validação do protótipo, como iremos observar nos capítulos quatro e cinco.

# 4 Implementação do protótipo

No decorrer do processo de melhoramento da ferramenta *data defender* constatou-se que uma empresa portuguesa, a RedGlue, estava igualmente a explorar as potencialidades da ferramenta e a desenvolver esforços para melhorar e adaptar a ferramenta para executar análises de dados pessoais, à qual designou de *ReDataSense* [50], publicitando e disponibilizando de uma forma aberta (através do *github*) as alterações efetuadas, explicando o racional das mesmas [51]. Algumas melhorias, por nós pensadas, já tinham sido efetuadas e implementadas parcialmente pela RedGlue, nomeadamente a inclusão da capacidade de reconhecimento de expressões regulares e direta de dicionários, não obstante e como se verificará neste capítulo, ainda houve margem para melhorar algumas das implementações promovidas.

Face ao exposto, o presente capítulo incide na explicação detalhada dos fatores que motivaram a realização das alterações efetuadas aos produtos referidos e de como foram implementadas. Para além dos melhoramentos efetuados na ferramenta (equivale à componente 1. Descoberta da arquitetura), é explicado como é que foram agregados os resultados (*output*) para servirem de entrada (*input*) numa nova componente edificada, a de governação e proteção (componente 2. e 3. da arquitetura), cujo resultado promoverá um relatório sumário (*dashboard*) do ponto de situação de um determinado repositório (componente 4. da arquitetura).

Como o objetivo principal do protótipo desenvolvido destina-se a realizar a descoberta de dados pessoais, o nome atribuído ao artefacto foi o de *PerDa2Disco* que deriva do inglês ***Per**sonal **Da**ta **to Disco**very*.

## 4.1 Criação de modelos NLP em português

Como referido, o ponto de partida foi a ferramenta de fonte aberta *data defender,* sendo que o autor disponibilizava um conjunto de modelos otimizados para a deteção de entidades na língua inglesa, recorrendo à utilização do modelo *Maximum Entropy* [25]. Uma vez que a ferramenta está desenvolvida na linguagem *Java*, para a construção de novos modelos, foi necessário recorrer aos utilitários de *OpenNLP* [26] e a scripts de anotação de textos para gerar os módulos de reconhecimento de dados pessoais para o referido modelo.

Neste sentido, o processo de anotação dos textos para treino dos modelos construídos, cuja criação foi baseada na adaptação de alguns códigos fonte disponibilizados nas páginas *web Apache OpenNLP* [26] e *tutorialKart* [52], nomeadamente para o treino de entidades nominais na língua portuguesa. Os documentos utilizados para a realização dos treinos foram baseados em coleções de livros escritos em português e de alguns artigos de imprensa.

Para além da identificação dos textos escritos em português foi necessário criar anotações nos textos de acordo com as entidades pretendidas. Para tal, numa primeira fase foi criado um script para permitir fazer as anotações de uma forma rápida e semiautomática. O script desenvolvido e utilizado foi o

comando *sed* [53] disponível nativamente nos ambientes *Linux,* cujo princípio de funcionamento é baseado na leitura do texto original escrito em português, comparando-o com a lista de entidades pretendida (por exemplo nomes). Sempre que uma entidade existente no dicionário seja coincidente com a do texto original, este introduz a anotação automaticamente em torno da entidade coincidente do texto, conforme a regra identificada na Figura 2.5 do capítulo 2, Trabalho relacionado (ver página 13). No final de executar o script, o texto passa a estar anotado e no final armazena o texto anotado com um novo nome, isto porque não se pretende alterar o texto original.

```
cat dicionarioNomes.txt | while read token
do
  if [[ $token != '' ]]
  then
    echo "$token"
    sed -ri "s/ ($token)([ .,;\!\?]))/ \<START:person\>\1\<END\>\2/g" nomeTextoAnotar.txt
  fi
done
```

Figura 4.1: Exemplo do script de anotação de textos

É importante referir que para a execução do script para gerar as anotações dos textos, ambos os ficheiros deverão ter o mesmo *encoding* de forma a garantir o reconhecimento integral das entidades. Quando existem entidades com acentuação é importante ter o cuidado de treinar a mesma entidade com letras maiúsculas e minúsculas para garantir uma melhor cobertura, garantindo que todos os termos foram devidamente anotados.

Depois dos textos estarem anotados, procedeu-se ao treino de reconhecimento de entidades nominais dos textos recorrendo à adaptação do exemplo do código disponibilizado pelo *tutorialKart* [52] e da página *web Apache OpenNLP* [26] para o treino de *NER* com a função *TokenNameFinder*. Essencialmente, o programa permite ler o documento de treino anotado e gerar os ficheiros binários de modo que a função *TokenNameFinder* permita ler e executar o ficheiro .bin gerado (por exemplo pt-ner-nome.bin).

Para o reconhecimento de palavras *(tokens)* e de frases *(sentences)* foram utilizados os módulos disponíveis na página da documentação da *Apache OpenNLP* [27].

Em resumo, o processo para a criação dos modelos de aprendizagem de língua natural foi realizado em quatro etapas:

1. Identificação e seleção de textos na língua portuguesa;
2. Criação de dicionários com as entidades a reconhecer;
3. Anotação dos textos com a etiqueta adequada.
4. Execução do modelo de treino para geração do ficheiro binário (.bin), para NER.

Embora tenham sido desenvolvidos esforços para criar módulos de ficheiros binários que permitam o reconhecimento de nomes, localidades, profissões, registos relacionados com a saúde e habilitações literárias, infelizmente os resultados obtidos não foram os esperados ficando muito aquém, principalmente pela dificuldade em conseguir identificar e recolher textos escritos em português

diversificados que contivessem o tipo de informação pretendida com um volume significativo de palavras mínimas para proceder ao treino. Apenas foi possível, fabricar o módulo para nomes, cujo treino foi baseado em várias obras literárias escritas na língua portuguesa, procurando abranger a totalidade dos nomes constantes nos dicionários criados.

Os maus resultados na criação de novos módulos binários, motivaram a criação de novos modos de descoberta que permitissem o reconhecimento de palavras por intermédio de consultas diretas aos dicionários e expressões regulares, como se poderá observar de seguida com a descoberta de dados.

## 4.2 Descoberta de dados (PerDa2Disco)

Como referido, a ferramenta *data defender* utiliza um conjunto de bibliotecas do *OpenNLP* para permitir fazer a segregação por palavras (*tokenizer)* e/ou por frases *(sentences),* paralelamente ao desenvolvimento dos vários módulos dos modelos binários criados na língua portuguesa (formato correto para permitir que a *data defender* consiga ler e interpretar), foram realizados alguns melhoramentos com o intuito de conseguir melhorar o processo de descoberta de dados pessoais.

Para o desenvolvimento do artefacto foi utilizado o ambiente *Windows* como Sistema Operativo (SO), recorrendo ao *NetBeans IDE 8.2*, cujas características são de acordo com a Figura 4.2, abaixo.

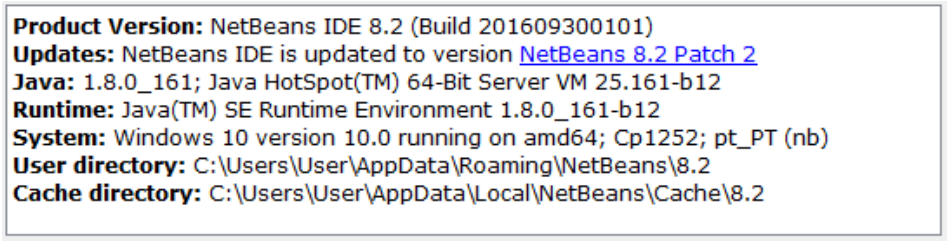



Figura 4.2: Características do IDE e SO

Foi igualmente necessário importar um vasto número de bibliotecas da *Apache OpenNLP* [54]*,* em virtude da exploração das ferramentas *OpenNLP* [55] estar dependente das mesmas, associando ao ficheiro *pom.xml* essas mesmas dependências, conforme é ilustrado na Figura 4.3.

```
<dependency>
    <groupId>org.apache.opennlp</groupId>
    <artifactId>opennlp-tools</artifactId>
    <version>1.8.1</version>
</dependency>
```

Figura 4.3: Dependências da ferramenta *OpenNLP*

Assim, seguidamente serão explicadas as principais melhorias realizadas no artefacto e os motivos que conduziram à sua realização.

### 4.2.1 Criação de novos modos de descoberta

Como já referido, a ferramenta *data defender* apenas permite executar a análise da descoberta de dados pessoais por intermédio dos modelos binários (ficheiros .bin) forçando, obrigatoriamente, a treinar modelos referentes a uma determinada entidade. A razão da limitação deve-se ao facto da ferramenta, em relação a este ponto específico, apenas ter configurado as bibliotecas de *tokenizerME* e *TokennizerModel* [25]. Aquando do processo da criação dos dicionários na língua portuguesa e das expressões regulares para reconhecer dados específicos portugueses, surgiu a ideia de incorporar a capacidade de analisar os documentos diretamente por intermédio dos dicionários e expressões regulares, motivo pelo qual foram exploradas e incorporadas novas bibliotecas possíveis de serem utilizadas. Concretamente, bibliotecas de leitura de dicionários *(Dictionary)*, de reconhecimento de nomes por dicionários *(DictionaryNameFinder)* e de reconhecimento de expressões regulares *(RegexNameFinder)*.

Neste sentido, foram alterados os dicionários para o formato *XML* e construídas várias expressões regulares, incorporando na ferramenta a capacidade de leitura direta dos dicionários e de expressões regulares, com recurso a um processo similar à leitura existente para os ficheiros .bin, através da função *NameFinder*.

Esta reestruturação do código fonte motivou a que na ferramenta, para além de se fazer o *tokenizer* com recurso apenas à biblioteca *NameFinder*, fossem incluídos novos casos para permitir o *tokenizer* para utilização dos dicionários e das expressões regulares. A designação atribuída para estes novos modos de descoberta foram *NERDictionary* e *NERRegex*, respetivamente. Após a inclusão destas novas funcionalidades, verificou-se que os resultados da descoberta de dados pessoais melhoraram substancialmente, para além de apresentarem um desempenho superior, melhoria de qualidade dos dados foi bastante significativa, como poderemos ter oportunidade de verificar no capítulo da avaliação.

Posteriormente, e após verificar a limitação do reconhecimento dos termos por palavra e não se estar a conseguir o reconhecimento de termos compostos complexos, devido à utilização de contrações de proposição e artigos definidos entre os termos, foi explorada a possibilidade de incluir a leitura caractere a caractere para além da leitura por palavra, criando um novo modo de reconhecimento de termos, designado por *NERPattern* porque se baseia na comparação de expressões padronizadas. Para o último modo foi necessário incluir novas bibliotecas, destacando-se a de *io.InputStreamReader,* de entre outras.

### 4.2.2 Descoberta baseada em múltiplos modos

Embora se tenham criado novos modos de descoberta, verificou-se a limitação de ser necessário executar mais do que uma vez a aplicação para os diferentes modos, isto é, ora se executa com a aplicabilidade de dicionários, ora com expressões regulares e assim por diante. Ao nível de funcionamento da ferramenta e com o intuito de otimizar a sua utilização, foi desenvolvida a possibilidade de explorar de uma forma integrada os vários modelos em simultâneo. O principal objetivo desta funcionalidade consiste em acelerar o processo de descoberta de dados pessoais, recorrendo a

diferentes modelos e alargando o espectro de reconhecimentos das entidades, assim como permitir simplificar o processo de classificação dos dados. Ao escolher mais do que um modelo a executar, com apenas uma execução a ferramenta está habilitada a avaliar os vários modelos parametrizados, devolvendo o melhor resultado (de acordo com a ponderação dada no mecanismo de procura) e para o caso de ter sido descoberta a mesma entidade, no mesmo local do documento por dois ou mais modelos diferentes.

Para tal, foi necessário atribuir fatores de ponderação (pesos) aos diferentes modos de descoberta, sendo que para (1) as expressões regulares foi atribuído um peso de 0,99 porque as regras traduzem algo muito específico e concreto, se a regra identificou é porque o termo identificado respeita garantidamente a regra; para (2) os dicionários foi atribuído um peso de 0,95 para permitir diferenciar as expressões regulares, mas o valor atribuído é igualmente elevado porque o reconhecimento das palavras baseia-se em dicionários (palavras especificas) e não em probabilidades, ou seja quando localiza uma palavra é porque a palavra existe, já para (3) o modo binário (MaxEnt) é de acordo com a probabilidade de um determinado dado corresponder a uma determinada classificação previamente treinada. Todos os modelos anteriormente referidos fazem uma análise por intermédio de *token* e como é possível ter um denominador comum que consiste na posição do *token* em relação ao documento, a ferramenta depois de analisar todos os módulos dos diferentes modos de descoberta vai comparar os vários termos, no caso de serem exatamente iguais e estarem na mesma posição relativa, devolve apenas o termo que apresente o peso maior.

Por exemplo, o modo MaxEnt descobriu o nome Ana na posição 110 (com uma probabilidade de 0,70) e o modelo dicionário descobriu o nome Ana na posição 110 (corresponde a um peso de 0,95) e o nome Flores na posição 111. Isto significa que o modelo MaxEnt não reconheceu a entidade Flores como sendo um nome, no entanto o modelo dicionário reconheceu o nome Flores, logo o resultado a devolver será Ana Flores. Neste caso, o resultado é influenciado pela ponderação do fator do modelo dicionário ser superior ao fator do modelo MaxEnt. Em situações que ocorra uma situação similar, mas que seja o modelo MaxEnt a detetar mais do que um *token* sucessivo em vez de ser o modelo dicionário, a escolha do resultado a devolver será o do modelo MaxEnt.

Uma das vantagens em executar simultaneamente mais do que um modelo consiste na redução do esforço em preparar o conjunto de dados (*dataset)* dos resultados *(logs)* para posterior análise e classificação dos dados pessoais em virtude de se eliminar à partida termos repetidos. Outra vantagem é que são complementares entre si, pois poderão existir termos que não constam no dicionário, mas se o modelo NLP estiver bem treinado poderá conseguir interpretar um determinado termo pelo sentido da frase, mesmo que esse mesmo termo não tenha sido utilizado para o treino. Neste caso, a probabilidade será reduzida, mas surgirá para posterior análise e interpretação do *dataset* gerado.

Finalmente, (4) o modo padrão foi considerado atribuir um peso de 0,98, superior ao valor atribuído aos dicionários porque os termos a procurar são termos compostos (constituídos por mais do que uma palavra ou por regras de padrões de acordo com expressões previamente definidas [56]), implicando um maior rigor do termo que é reconhecido. No entanto, ao contrário dos modos de descoberta anteriores a sua análise é baseada caractere a caractere o que implica que não exista um denominador

comum para permitir relacionar os termos descobertos. A desvantagem prende-se com o facto de quando se executam dois módulos idênticos (por exemplo nomes) de modos de descoberta diferentes, os resultados obtidos serão duplicados dificultando a sua análise.

### 4.2.3 Dicionários

Embora os dicionários tenham sido utilizados inicialmente para permitirem desenvolver novos modelos de NER para a língua portuguesa, acabaram por ser mantidos como modo de pesquisa da ferramenta. A alteração que foi necessária realizar cingiu-se a converter os dicionários para o formato *XML* e promover as alterações elencadas anteriormente aquando da criação dos novos modelos.

No total foram criados e/ou melhorados dez dicionários diferentes considerados relevantes para o tema em questão e que de alguma forma o termo utilizado no dicionário tenha uma forte relação com o tipo de dado pessoal que se pretende descobrir, para poder identificar univocamente um determinado titular dos dados. Os dicionários que estão disponíveis a ser utilizados são:

1. Crimes. Conforme a tipificação dos crimes do Código Penal [57];
2. Localidade do Código Postal. Com base na informação oficial dos CTT [58];
3. Estado Civil. Apenas inclui os estados reconhecidos pelas finanças [59];
4. Filiações. Várias fontes da internet;
5. Género. Várias fontes da internet [60];
6. Habilitações literárias. De acordo com a lista da Direção-Geral da Administração e do Emprego Público [61];
7. Nomes. Apenas são apresentados os nomes reconhecidos e aceites pelo governo português [62];
8. Profissões. Diversas fontes da internet;
9. Religiões. As mais comuns;
10. Dados relacionados com saúde. Todos os registos utilizados foram retirados do site do Ministério da Saúde [63] [64].

A utilização de dicionários foi um bom exercício para sentir as dificuldades existentes no NLP, principalmente com as ambiguidades de determinados termos muito utilizados na língua portuguesa e que poderão ter mais do que um sentido. Embora não seja a forma mais correta, procurou-se reduzir algumas ambiguidades prevalecendo determinados termos de um dicionário em detrimento de outros, mas ainda assim foram muito poucos os termos que se optou por esta forma de gerir os conflitos com as ambiguidades. Para além de se considerar que o tratamento das ambiguidades deverá ter uma abordagem mais inteligente e pragmática, como se poderá verificar no capítulo das conclusões, no trabalho futuro, considerou-se que é preferível descobrir um determinado termo mesmo que a contextualização não seja a mais correta, do que não descobrir e poder ser relevante para o estudo, isto porque um dos objetivos primários da ferramenta é que funcione como um despertador de alertas para a existência de potenciais dados pessoais.

### 4.2.4 Expressões regulares com validadores

Com o objetivo de permitir ter uma melhor flexibilidade e escalonamento das regras pretendidas, foi adaptado o código fonte para permitir acomodar as expressões regulares segregadas do código fonte e ter a capacidade de ler um ficheiro de propriedades independente. Para esta funcionalidade recorreu-se ao utilitário *Regex* do *opennlp.tools.namefind.RegexNameFinder* [65] e foram incorporados métodos de validação no próprio código fonte, para o caso de ser detetado uma determinada expressão regular.

A grande vantagem da inclusão das expressões regulares com esta configuração é permitir a modularidade e escalabilidade para se criar expressões regulares de acordo com o padrão que se pretende. O único cuidado a ter é que deve respeitar as regras *regex* para o tipo de linguagem Java. Para a formulação das várias expressões regulares recorreu-se a sites específicos para este tipo de testes [66] e a vários tutoriais existentes para o efeito [67].

Para o desenvolvimento do protótipo o ficheiro de propriedades *Regex* disponibiliza, atualmente, várias expressões regulares das quais se destacam as seguintes:

1. Cartão Cidadão (CC) [68];
2. Número de Identificação Fiscal (NIF) [69];
3. Número de Identificação de Segurança Social (NISS) [70] [69];
4. Endereços de correio eletrónico [71];
5. Código-Postal apenas compostos com 4+3 digitos;
6. Telemóveis e telefones fixos sem o número internacional;
7. Carta de condução portuguesa [72];
8. Cartões de crédito do tipo VISA, *Master Card*, *American Express*, *Diners*, *Discover*, JBC, *Maestro* e *Payment Card* [69];
9. Número de Identificação Bancária (NIB) e IBAN [69] [73];
10. Datas abreviadas no formato europeu, da norma ISO 8601 [74] e Norma Portuguesa[17].

A regra para a identificação do cartão cidadão português foi reajustada para permitir o reconhecimento do número (i) com apenas os oito dígitos, sem o digito de validação; (ii) com nove dígitos onde se inclui o digito de validação antigo (o utilizado no bilhete de identidade antes da entrada em funcionamento do cartão do cidadão); e (iii) com 12 dígitos, inclui o digito de validação (nono algarismo) e o conjunto dos três novos dígitos de validação introduzidos com o cartão do cidadão. Neste último caso, o número poderá estar representado com os algarismos todos seguidos ou agrupados em dois blocos separados com hífen, sendo o primeiro bloco o número de oito dígitos e o segundo bloco os dígitos de controlo com quatro algarismos. De modo a se ter um melhor entendimento, seguidamente são apresentados alguns exemplos dos formatos que a expressão regular permite reconhecer:

| | |
|---|---|
| 12345678 | (i) |
| 123456789 | (ii) |
| 123456789ZZ1 | (iii) |

[17] NP EN 28601

12345678-9ZZ1 (iii)

Adicionalmente à regra, foi introduzido um validador do documento do cartão do cidadão [68]. No entanto, a verificação do número pelo validador está limitada ao número composto pelos 12 dígitos. Isso deve-se ao facto de serem os últimos quatro dígitos do cartão do cidadão responsáveis pela validação do número [75].

No que concerne ao Número de Identificação Fiscal (NIF), a expressão regular foi reajustada para reconhecer apenas os números que digam respeito a particulares, ou seja, os números começados pelo algarismo um ou dois. Fruto da sua modularidade, rapidamente se poderá incluir o reconhecimento dos números iniciados pelo algarismo três, pois este está reservado para os particulares aquando se esgotar os começados por dois.

De modo a incrementar o rigor na deteção do tipo de dados que possam identificar univocamente um determinado indivíduo (de acordo com a Tabela 3.1), foram acrescentados no código fonte diferentes validadores. Para além do Cartão de Cidadão, anteriormente referido, foram criados validadores para o NIF [76], o NISS, cartões de crédito e NIB [69]

### 4.2.5 Reconhecimento de termos compostos

Outro melhoramento foi a introdução da capacidade no reconhecimento de termos compostos, ou seja, a junção de mais do que um termo consecutivo, com o intuito de incrementar o rigor dos dados pessoais descobertos, permitindo, por exemplo, no caso da ferramenta, descobrir os termos Maria e José, se os dois termos estiverem seguidos um do outro. É pretendido que o resultado a retornar seja apenas uma entidade com o nome de Maria José e não duas entidades, uma Maria e a outra José. Já no caso dos termos não sejam seguidos, o resultado deverá ser duas entidades distintas.

A funcionalidade acima referida apenas é aplicada nos modos de descoberta MaxEnt e Dicionários.

O mecanismo utilizado é baseado na composição dos termos existentes nos dicionários e o nos módulos desenvolvidos através de treinos de texto (ficheiros binários) e na leitura dos documentos a analisar. Isto significa que no decorrer da análise aos documentos, sempre que surja uma ou mais palavras (*token*) sequências que constem nos módulos referidos, irão ser interpretadas como pertencentes ao mesmo grupo de termos.

A grande vantagem desta pequena funcionalidade é que permite reduzir o número de termos reconhecidos de um determinado grupo de dados pessoais, nomeadamente quando se está perante a termos compostos. Por exemplo, se num determinado documento constar uma frase do tipo “olá, eu sou o Rui Santos e resido em Almada.” Através da utilização do módulo “Nome”, do modo de descoberta dicionário e MaxEnt a ferramenta devolverá como resultado o termo Rui Santos como sendo um nome. Infelizmente, não foi possível incluir o tratamento dos termos quando apresentam contrações da preposição (“do”, “de”, “da”, …) ou a conjunção coordenativa (“e”) ou outro qualquer artigo definido do tipo (“a”, “o”), entre dois termos existentes a ferramenta não conseguirá interpretar o termo como sendo o mesmo, mas sim indicará como sendo dois termos diferentes. Por exemplo, um documento que

conste os nomes “Figueiredo dos Santos” e “Pereira de Almeida” será devolvido no total quatro nomes diferentes, em vez de apenas dois nomes.

### 4.2.6 Consulta personalizada

Foi adicionada a capacidade de realizar consultas através de padrões personalizados onde se incluem pesquisas por termos compostos. Neste caso, ao contrário dos restantes modelos, a consulta por padrões não é realizada por intermédio de *tokens* mas sim através de caractere a caractere. O principal motivo que originou o desenvolvimento desta funcionalidade consistiu em permitir particularizar uma determinada consulta, ou seja, procurar um conjunto de dados específicos relativos a uma determinada pessoa. Essencialmente, foi a pensar numa situação muito particular do RGPD, de uma possível reclamação por parte de um titular de dados a querer saber como e onde é que os seus dados pessoais estão armazenados. Com esta funcionalidade, é possível fazer um varrimento à rede ou procurar especificamente em determinados repositórios os dados concretos de apenas um determinado titular, sobre diferentes formas de edição. Por exemplo, na consulta poderemos definir o nome da rua do titular, de uma forma completa ou abreviada (“Rua da Pecuária” ou “R. da Pecuária” ou apenas “R Pecuária”, …), à semelhança do nome da rua poderá ser definido outro tipo de dados que possam estar associados ao titular de dados.

Paralelamente ao motivo principal, foi pensado explorar este modelo para permitir aferir com maior rigor os resultados encontrados, nomeadamente os termos compostos, tais como um endereço de uma morada, nome completo ou outro qualquer tipo de procura formada por mais do que um termo composto que inclua uma contração, conjugação coordenativa ou artigo definido. Isto foi possível por o modelo utilizar igualmente utilitários baseados em expressões regulares, mais concretamente o *java.uitl.regex* funcionando numa base de compilação e verificação de padrões [56].

A desvantagem da utilização por padrões, como se poderá observar no capítulo 5. Avaliação, é que torna a pesquisa mais lenta e demorada, principalmente quando o volume de documentos é elevado e por apresentar ocorrências dos registos duplos, esta situação acontece quando o modo identifica, por exemplo o termo “Pereira de Almeida”, ele não consegue entender que o termo “de Almeida” se refere ao mesmo termo completo, isto é resolve dois termos: (1) “Pereira de Almeida” e (2) “de Almeida”. Outra situação possível consiste quando é identificado um primeiro termo antes do termo que tem a contração da preposição, por exemplo, caso reconheça “Raúl Pereira de Almeida”, irá igualmente identificar o termo “Pereira de Almeida” e o “de Almeida”, neste caso apresentaria três registos em vez de um. Esta situação ocorre em virtude de a ferramenta iniciar a comparação de termos com base no primeiro caractere descoberto, e como deu para perceber, nos exemplos dados o número da posição do primeiro caractere em relação ao documento é diferente.

Por esta razão, para a realização de uma análise de dados num repositório de dados de dimensões significativas, não se recomenda iniciar o processo de descoberta de dados com a inclusão deste modelo, ou seja, a utilização deste modelo deverá ser utilizada para quando se sabe que tipo de dados se pretende descobrir. Assim, no caso de se utilizar este modo de leitura deve-se ter atenção para não se utilizarem os módulos correspondentes aos modos de descoberta por dicionário e MaxEnt.

### 4.2.7 Procura recursiva local ou em rede

Na descoberta de dados pessoais em dados não estruturados, foi melhorado o modo de pesquisa recursiva num determinado repositório ou numa determinada estrutura de pastas existentes localmente num computador ou numa rede de computadores. A parametrização poderá ser feita escrevendo (1) simplesmente o nome da pasta; (2) mapeamento do endereço de rede e/ou (3) o endereço do IP pretendido. Em qualquer das situações, para além de realizar a procura na pasta ou endereço de rede, irá procurar recursivamente por todas as sub-pastas existentes na pasta/endereço de rede definido. No caso de ser configurada uma procura através de um mapeamento de rede ou IP, deverá ter obrigatoriamente permissões de acesso. O objetivo desta funcionalidade consiste em poder executar a ferramenta remotamente, de uma forma centralizada pelo responsável dos sistemas de informação, sempre em coordenação com os administradores de rede, para salvaguardar e minimizar uma posição violação na privacidade de dados pessoais.

### 4.2.8 Pesquisa por amostragem aleatória

Foi alterado o modo como inicia o processo de descoberta de dados pessoais. Em vez de realizar a procura por omissão em todos os tipos de ficheiros com exclusão apenas de alguns, por exemplo ficheiros do tipo *dll*, foi alterado para realizar a procura por inclusão dos ficheiros. A razão que motivou esta alteração consistiu em permitir otimizar o desempenho, quer na localização, quer na análise, através da parametrização de formatos previamente conhecidos cujo seu conteúdo, tipicamente, é do género de texto. Desta forma, o utilizador é obrigado definir deliberadamente o tipo de ficheiro a pesquisar e poderá ser utilizado mais do que um tipo de extensão, devendo os tipos de extensão estar separados por vírgula sem espaço. Ver exemplo da Figura 4.4, abaixo:

```
# Extensão do(s) ficheiro(s) a descobrir e analisar (separar por vírgula sem espaço)
inclusions=docx,doc,docm,xlsx,xls,pdf,txt,pptx,ppt
```

Figura 4.4: Parametrização dos ficheiros a analisar

Foi alterado o script de execução da ferramenta para permitir ser executada de uma forma parcial, ou seja, quando se executa a ferramenta. Inicialmente irá ser efetuado um varrimento ao repositório/diretório a todos os ficheiros encontrados com as extensões definidas, como referido anteriormente, este varrimento irá ser realizado em todas as pastas e subpastas existentes, desde que existam ficheiros com o tipo de extensão. O resultado será uma listagem de todos os ficheiros existentes com a respetiva localização, assim como o total dos ficheiros existentes. Culminado com a questão se pretende analisar os ficheiros supracitados? (y/n).

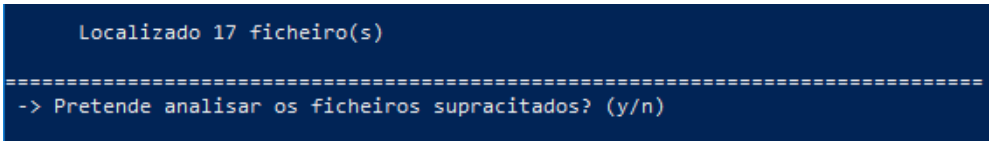



Figura 4.5: Informação dos documentos encontrados

Caso se pretenda executar a análise para a descoberta de dados pessoais, os documentos serão analisados sempre de um modo aleatório, independentemente do estudo ser realizado por censo ou por amostragem. No caso do estudo ser efetuado por censo (totalidade dos elementos da população) não seria necessário realizar uma análise de uma forma aleatória, já para o caso do estudo ser por amostragem (pequena parte de todos os elementos da população), como se pretende que a avaliação seja válida e representativa do todo, considera-se que o método mais adequado seja por intermédio de uma seleção aleatória dos documentos para evitar tendências de manipulação de resultados.

Para a elaboração de um plano por amostragem é preciso definir o tamanho da amostra, motivo pelo qual que para além de ter sido introduzido a aleatoriedade na análise dos ficheiros, foi igualmente introduzido um limitador para o caso de o repositório possuir um volume significativo de documentos (ficheiros) e se pretender realizar a descoberta da existência de dados pessoais por amostragem. Caso o limitador seja superior aos ficheiros localizados irá analisar todos os ficheiros localizados.

```
# Limitar a procura dos ficheiros encontrados
limit=100
```

Figura 4.6: Parametrização do limitador

Outra funcionalidade que foi adicionada, consiste na possibilidade de se poder excluir um ou mais ficheiros específicos, não requer saber a localização exata do ficheiro, basta apenas conhecer o nome do ficheiro. À semelhança das outras funcionalidades, este campo poderá ser parametrizado com mais do que um ficheiro, devendo os mesmos estarem separados por vírgula e sem espaço. Esta necessidade, surgiu durante a fase de testes de avaliação que por determinados momentos não se pretendia analisar um ou outro documento e para não estar sempre a remover o documento da pasta foi incluído esta modesta funcionalidade.

```
# Caso se pretenda excluir um ou mais ficheiro especifico
files_excluded=a.txt
```

Figura 4.7: Parametrização dos ficheiros a excluir

### 4.2.9 Automatização da classificação

De modo a automatizar o processo da classificação dos dados pessoais, foi introduzida uma classificação de grau de risco de acordo com a matriz de risco identificada na Tabela 2 (ver página 27, seção 3.3, Modelo proposto). Assim como, para além do registo de relatórios (*logs*) de informação (*info*), de despiste do funcionamento da ferramenta (*debug*) e eventuais erros que possam ocorrer (*error*), os resultados para posterior análise e governação são agregados em três ficheiros diferentes cuja extensão é do tipo .csv, de modo a permitir diferentes tipos de análises:

a) Resultado. Pretende devolver todos os resultados encontrados juntamente com alguma informação mais detalhada, para uma eventual análise técnica e/ou poder correlacionar diferentes tipos de dados a um único titular;

b) Resumo. É agregado de acordo com o nome de ficheiro para permitir avaliar a densidade de risco de um determinado repositório, independentemente de ser estruturado (base de dados) ou não estruturado (serviço de diretório);
c) Governação. Com o objetivo de poder auxiliar do EPD de uma determinada organização, a ter uma visualização mais abrangente das dependências existentes entre os vários documentos e se for o caso da dependência entre repositórios permitindo carregar esta informação numa outra ferramenta de visualização por grafos, simplificando a leitura e interpretação do volume de dados pessoais existentes.

Para o cálculo da densidade do risco, foram incorporados ao nível do código todas as fórmulas de cálculo e adicionado um novo ficheiro de propriedades de modo a permitir a parametrização do fator de ponderação do tamanho dos documentos (ver seção 3.3 Modelo proposto, página 31).

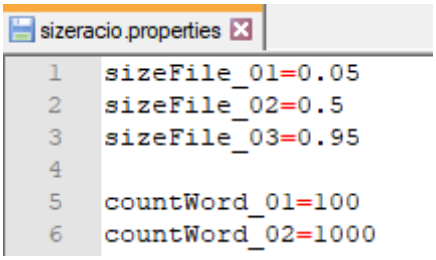


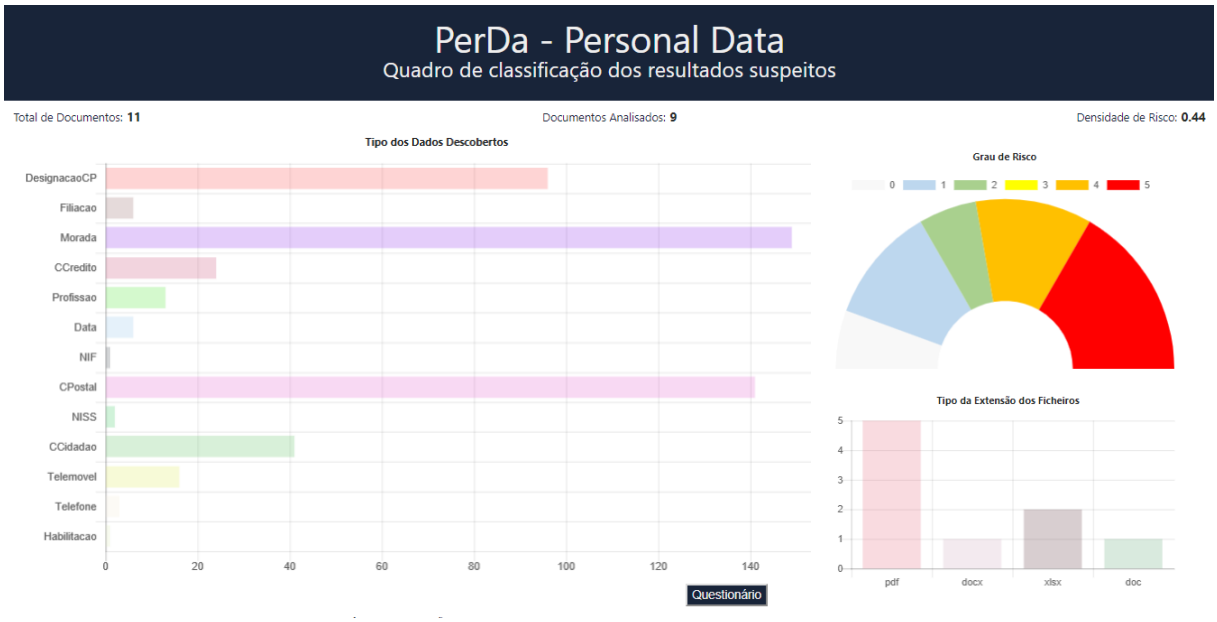

Figura 4.8: Rácio do tamanho dos documentos

### 4.2.10Apresentação dos resultados

Embora seja necessário recorrer à linha de comandos para colocar em funcionamento o protótipo, foi desenvolvida uma interface gráfica em HTML, recorrendo-se à linguagem de *Javascript*, utilizando a estrutura de trabalho da W3.CSS [77] em virtude de ser de fonte aberta e apresentar uma boa capacidade de resposta integrada, sendo muito simples e de fácil desenvolvimento e bibliotecas da jquery.mim.js, Chart.min.js e Chart.bunble.min.js para a apresentação dos vários gráficos desenvolvidos permitindo uma melhor visualização dos resultados encontrados [77].

Os resultados dos dados descobertos, no caso dos dados não estruturados, foram agrupados de forma a permitir ter uma perceção do volume total de documentos existentes em relação à amostra dos dados analisados, a densidade de risco num determinado repositório e os tempos de localização e de execução. Já para os dados estruturados, a ferramenta indica o volume total de registos existentes (total de linhas existentes entre as várias tabelas) em relação aos registos da amostra, a densidade de risco e o tempo de execução.

Ilustra graficamente, a classificação do tipo dos dados pessoais descobertos permitindo ter uma visão global da dimensão dos vários tipos de entidades (lado esquerdo através de um gráfico de barras horizontal), assim como a classificação de risco (gráfico semicircular no lado direito em cima), de acordo com os critérios de classificação da criticidade do risco proposta (ver seção 3.3, Modelo proposto, página 27) e um pequeno gráfico de barras vertical, do lado direito, em baixo, com a agregação do formato dos documentos e/ou área de negócio.




| Tipo Ficheiro | Pasta | Ficheiro | Tipo de Dado | Clas. Risco | Densidade Risco | Nr Dados | Tamanho Documento |
|---|---|---|---|---|---|---|---|
| pdf | D:\TESTES\teste\ | 20180402_MASTERCLASS-MARCIA_75.pdf | Filiacao, IBAN, CPostal, Data, DesignacaoCP, Nome | 5 | 0,05556 | 17 | 235 |
| docx | D:\TESTES\teste\doc\ | teste.docx | Telemovel, NIF, Nome | 4 | 0,04444 | 7 | 214 |
| pdf | D:\TESTES\teste\ | Teste2_CI.pdf | Filiacao, Profissao, Data, Nome | 5 | 0,11000 | 17 | 2460 |

Figura 4.9: Aspeto gráfico do resultado da ferramenta

Finalmente, na parte inferior, lista em forma de tabela todos os documentos analisados com a indicação da localização do ficheiro no repositório (caminho), o nome do ficheiro, o tipo de entidades descobertas, qual a classificação de risco associado ao documento, a densidade de risco parcial, o número de entidades identificadas no documento e a dimensão do documento em palavras.

Imediatamente, abaixo do gráfico de barras horizontal, no seu canto inferior direito encontra-se um botão com a descrição do questionário para dar seguimento à avaliação de segurança sobre os mecanismos utilizados para a proteção de dados pessoais.

Os resultados da produção gráfica resultam da agregação dos resultados recorrendo aos ficheiros de registos *(logs)* no formato .csv, um com o nome “Resultados” e o outro com o nome “Resumo”.

## 4.3 Governação e proteção

Como referido no capítulo 3. Solução proposta, pretende-se com a componente de governação e proteção de dados ter uma visibilidade da relação dos dados pessoais e conhecer em que medida é que estão a ser implementados e monitorizados os mecanismos de segurança tecnológicos em relação à privacidade de dados pessoais. Neste sentido, os recursos utilizados para o desenvolvimento do artefacto consistiram em *Javascript,* utilizando a mesma estrutura de trabalho da W3.CSS [77], para a

formulação e visualização do modelo do plano de mitigação apresentado (ver Figura 3.4, página 34) as possíveis ameaças no sector tecnológico foram traduzidas num questionário de avaliação e a obtenção dos resultados com o plano de mitigação.

De forma a agilizar o processo de avaliação de acordo com a classificação dos dados pessoais descobertos, considerou-se mais adequado que a avaliação fosse efetuada no seguimento da descoberta de dados, motivo pelo qual foi incluído o botão para se avançar para o questionário.

A resposta a todas as questões é de caracter obrigatório, caso não se responda a uma ou mais questão e se tente submeter o questionário surgirá uma indicação de: “Atenção!!! Resposta a todas as questões…” passando para cor vermelho todas as questões que ficaram por responder.

Sempre que se selecionar a opção de “não”, a uma determinada questão, o relatório de avaliação final indicará quais são as questões que não estão conformes, sugerindo um plano de ação para mitigação e/ou resolução da(s) ameaça(s). Após responder ao questionário, será promovido um quadro final de avaliação com as medidas tecnológicas que deverá adotar, para garantir uma cultura de privacidade de dados pessoais, reforçando novamente a classificação do tipo de dados existentes num determinado repositório, a classificação de densidade de risco e adicionando o fator risco. Em complemento ao plano de mitigação sugerido, o utilizador poderá fazer, se assim o desejar, uma avaliação de impacto sobre os aspetos não conformes, respondendo às ameaças/vulnerabilidade de acordo com a Tabela 3.2: Impacto do evento e a Tabela 3.3: Probabilidade de ocorrência (ver página 29), para ter uma indicação do nível de gravidade que a não resolução da medida virem a ter (ver Figura 3.2: Matriz de Probabilidade versus Impacto, ver página 30).

Embora o artefacto desenvolvido produza como resultado (*output*) um ficheiro com o nome Governação no formato .csv o seu carregamento na ferramenta do EAPY terá de ser efetuado de forma manual, através da componente *back-Office* [45]. É possível carregar vários ficheiros e depois criar as dependências entre repositórios, no caso de existirem, pelo que embora o artefacto desenvolvido permita fazer pesquisas em mais do que um repositório em simultâneo será mais vantajoso executar o artefacto individualmente por repositório.

Após importar as várias consultas pretendidas (documentos), o campo *Overview* no menu da componente de *back-Office* permite ter uma visualização global dos repositórios analisados com as respetivas relações (ver Figura 4.10, abaixo). Para além do *back-Office*, a ferramenta está habilitada com um *front-Office* que apresenta um conjunto de menus simples e interativos de navegação, contudo para explorar devidamente a ferramenta na sua componente de *front-Office* é necessário que se definam as vistas dos dados (*Data Views*), nomeadamente através dos objetos criados (*Data Objects*) e das relações pretendidas (*Data Relation*) [45]. No limite, pode ser definida uma relação total entre todos os objetos, no entanto se se pretender realizar uma avaliação mais cuidada, as relações entre os vários objetos deverão ser criadas, recomendando-se que seja seguido o modelo de dados proposto, conforme a Figura 3.3 (ver página 33), caso contrário não é possível ter uma vista por intermédio de grafos das relações e dos objetos criados.

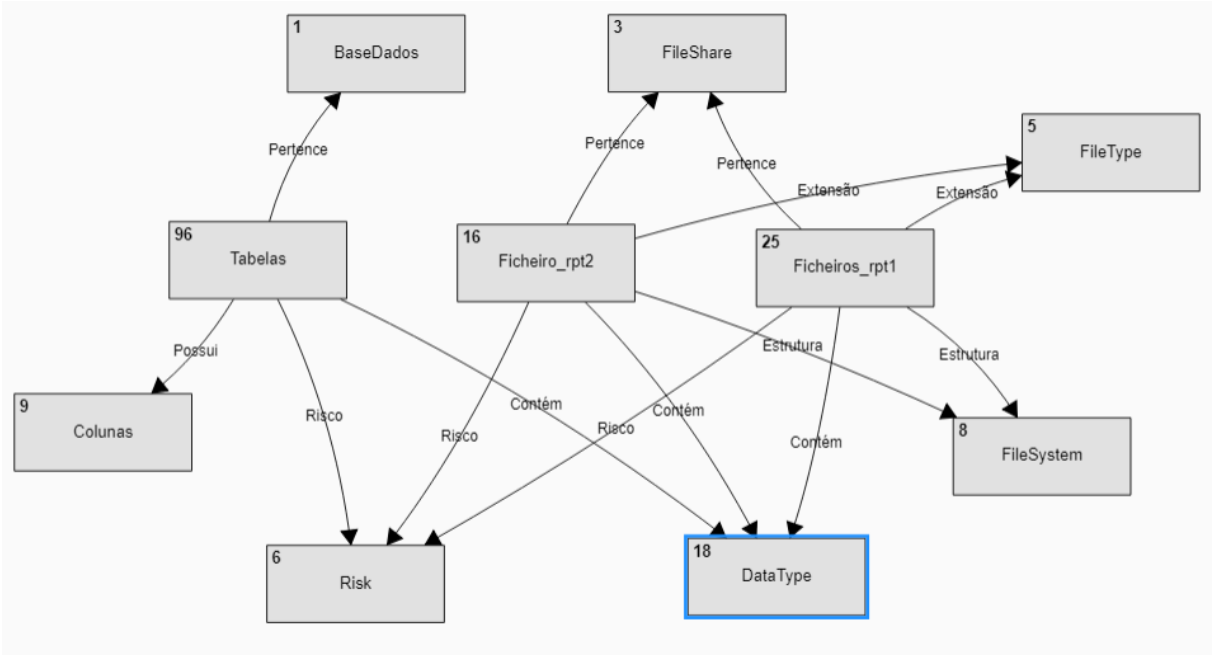



Figura 4.10: Vista de grafos do *back-Office* da ferramenta EAPY

## 4.4 Sumário

Como ponto de partida para o desenvolvimento do artefacto foi utilizada a ferramenta *data defender* com o objetivo de a tornar capaz de identificar e reconhecer dados pessoais na língua portuguesa de, uma forma otimizada e eficiente, promovendo uma classificação do tipo de dados de acordo com o RGPD, com uma abordagem para a sensibilização de medidas tecnológicas a fim de minimizar desvios na governação e na gestão da privacidade dos dados pessoais, passível de ser aplicada em dados estruturados e não estruturados.

Como se pode verificar, a implementação do artefacto foi segregada em três fases distintas: (1) criação e desenvolvimento de dicionários e expressões regulares para poder criar os módulos de tipos de dados de acordo com o modo *Maximum Entropy*, o que conduziu forçosamente à identificação de textos em português com diferentes atributos para permitir o treino dos referidos modelos; (2) melhoramento da ferramenta em si, adicionando a capacidade de ler novos modelos (dicionários, expressões regulares e padrões) para além do inicial *Maximum Entropy*, promovendo uma aproximação do reconhecimento de termos compostos e automatizando a classificação de dados pessoais; finalmente (3) edificação da capacidade inicial de governação dos dados pessoais, através da visualização gráfica dos resultados devidamente agregados, permitindo conhecer as relações existentes em relação ao repositório e apresentação de um conjunto de medidas de proteção, na vertente tecnológica, bem como, a possibilidade de se poder fazer uma avaliação de impacto sobre os dados pessoais descobertos.

O protótipo foi desenvolvido num ambiente *Windows*, utilizando a linguagem *Java* e o *NetBeans IDE* 8.2, recorrendo a um vasto conjunto de bibliotecas já existentes, permitindo simplificar a implementação do reconhecimento de entidades e a linguagem *Javascript* para as componentes de governação, proteção e relatório. Não obstante, toda a componente das anotações, quer dos textos de treino quer

dos textos para a realização de testes de laboratório foi desenvolvida em ambiente *Linux* (*Ubuntu LTS* 16.04), com a criação de scripts para anotar textos.

Embora não se tenham verificado constrangimentos durante a fase da implementação, nomeadamente dificuldades na operacionalização de alguns mecanismos que poderiam melhorar a deteção de dados pessoais, foi possível produzir um produto, cujas principais características e funcionalidades são:

- Disponível 4 modos de descoberta:
    - Expressões Regulares; Dicionários; MaxEnt e Padrões Específicos.
- Pesquisa de dados Estruturados e Não Estruturados;
- Modo de pesquisa integrada de mais do que um modelo;
- Pesquisa de dados em pastas/subpastas, localmente ou em rede;
- Procura por amostragem aleatória;
- Resultados em ficheiro .csv *(dataset);*
- Alerta de Densidade de Risco.

E capaz de identificar e reconhecer um alargado tipo de dados pessoais:

- Cartão do Cidadão;
- Número de Identificação Fiscal;
- Número de Identificação da Segurança Social;
- Endereços de Correio Eletrónico;
- Cartão/Licença de Condução;
- Telemóveis e Telefones fixos;
- Reconhecimento de Cartões de Crédito (PCI);
- IBAN / NIB;
- Nome, Género, Profissão, Habilitações Académicas;
- Morada, Código Postal;
- Registo Criminal (Crimes mais comuns);
- Doenças (Doenças comuns e algumas raras);
- Filiações (Sindicais, desportivas, organizações, …);
- Origem racial ou étnica.

# 5 Avaliação

A criação de novos módulos de NER a serem executados no artefacto, implica que sejam testados segundo duas perspetivas: de desempenho e de qualidade. Para ambos os casos foi definida uma metodologia de avaliação, identificando as métricas a utilizar e o conjunto de testes a executar. Numa primeira fase, os testes foram realizados em ambiente controlado (laboratório) através da criação de cenários com a existência de dados pessoais devidamente identificados, com o objetivo de analisar com detalhe os resultados mediante uma linha base referencial para posteriormente, poder aplicar os modelos criados em cenários próximos de um ambiente de produção (estudos de caso).

A motivação para a escolha dos estudos de caso, centrou-se em poder utilizar o artefacto em duas organizações com diferentes abordagens, mas que comungam da necessidade de tratar dados pessoais. Procurou-se assim, demonstrar as funcionalidades do protótipo e como é que o mesmo poderá ser vantajoso na ajuda de localizarem a existência de dados pessoais. Assim, foram efetuados dois estudos de caso: (1) na Marinha, através da procura de dados pessoais na secretaria partilhada do sector da Superintendência das Tecnologias de Informação (STI) que apresenta documentos repartidos entre servidor colaborativo da STI e os computadores atribuídos a colaboradores; e (2) em repositórios de dados do sistema *Edoclink* da empresa Link Consulting, por ser uma aplicação de gestão documental composta por documentos com muita informação de gestão. Por esse motivo o estudo de caso incluiu duas componentes distintas: (a) procura em repositórios de base de dados (dados estruturados), no caso concreto *SQLServer* da Microsoft, e (b) no servidor onde se encontram armazenados os documentos (dados não estruturados).

## 5.1 Laboratório

Nos testes em laboratório foram criados vários cenários para a descoberta de dados pessoais, cuja procura incidiu em dados não estruturados e dados estruturados. A criação dos vários cenários foi efetuada à medida que a ferramenta foi sendo desenvolvida e/ou melhorada para testar as várias funcionalidades pretendidas. Uma avaliação efetuada nos testes de laboratório consistiu na medição do nível de exatidão dos dados pessoais descobertos, através de métricas comummente utilizadas para a avaliação de reconhecimento de entidades em sistemas de língua natural: Precisão *(precision)*, cobertura *(recall)* e *F-measure* [20].

Para perceber a diferença entre as métricas de avaliação é importante ter em conta o conceito da matriz de confusão *(confusion matrix)*. Na prática, este conceito consiste numa tabela usada para descrever o desempenho de um modelo de classificação, face a um determinado conjunto de dados, para os quais são conhecidos os valores reais.

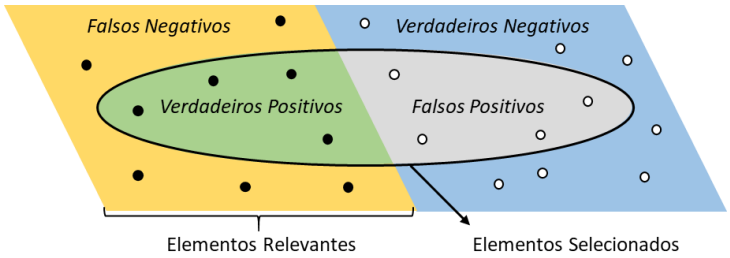



Figura 5.1: Ilustração gráfica da matriz de confusão. Adaptado de [78]

Com recurso ao ilustrado na Figura 5.1, foram deduzidas as fórmulas a seguir expressas para realizar a avaliação do cenário concreto da descoberta de dados pessoais, de acordo com:

- Precisão: baseia-se na razão das observações positivas corretas pelo total observações positivas obtidas. Uma precisão elevada significa essencialmente que estamos perante um número reduzido de falsos positivos.

$$Precisão = \frac{VP}{VP + FP\ (6)} \quad (4)$$

- Cobertura: também conhecida como *sensitivity*, em virtude de medir a sensibilidade do modelo, isto é, a percentagem de palavras relevantes que foram descobertas. Considera-se que para valores acima dos 50% é um bom modelo.

$$Cobertura = \frac{VP}{VP + FN\ (7)} \quad (5)$$

  Onde:

  – VP (Verdadeiros Positivos) = Observações Selecionadas Relevantes Corretas.
  Incluem as observações que mesmo não sendo totalmente iguais apresentam um forte indicador de similaridade com o termo correto;
  – FP (Falsos Positivos) = Observações Selecionadas não Relevantes.
  Inclui as observações com termos ambíguos *(OA)*, poderia ser, mas está fora do contexto; termos duplicados *(OD)* e termos errados *(OE)*, que por si só não é possível afirmar que seja um dado pessoal;

$$FP = OA + OD + OE \quad (6)$$

  – FN (Falsos Negativos) = Observações não Selecionadas Relevantes.
  Fazem parte deste grupo as observações omissas *(OO)* que deveriam ser detetadas, mas não o modelo não foi capaz de identificar.

$$FN = OO \quad (7)$$

- Finalmente, como medida de ponderação entre a precisão e cobertura, temos o *F-measure* [20] que é útil quando se está perante uma distribuição de classes desiguais:

$$F - measure = \frac{2 * Precisão\ (4) * Cobertura\ (5)}{Precisão\ (4) + Cobertura\ (5)} \quad (8)$$

Para a avaliação de qualidade foram identificados 30 documentos com diversas informações pessoais, de diferentes fontes e de diferentes formatos (faturas, *curriculum vitae*, requerimentos, declarações, entre outros documentos). Todos os documentos foram revistos manualmente de forma a confirmar e marcar (etiquetar) o tipo de dado pessoal existente, para permitir calcular a relação entre as entidades nominais descobertas e as reais. Em conjunto, os documentos apresentam 5 890 parágrafos e 10 178 tipos de dados pessoais, distribuídos de acordo com a Tabela 5.1.

O teste consistiu em executar o protótipo com recurso aos vários módulos criados para os diferentes modos de descoberta sobre os 30 documentos anteriormente referidos. Posteriormente foram comparados manualmente os resultados obtidos da descoberta e classificação dos dados pessoais em relação aos documentos previamente conhecidos aferindo a veracidade dos mesmos. O processo consistiu em verificar manualmente todos os registos dos dados descobertos marcando-os de acordo com as métricas anteriormente referidas *(VP, AO, OD, OE e OO)* permitindo realizar os respetivos cálculos.

Os resultados apresentados na Tabela 5.1, abaixo, são fruto dos melhores resultados conseguidos entre os vários módulos criados e disponíveis.

Tabela 5.1: Resultados das métricas de qualidade

| **Tipo de Dado**[18] | **Qtd**[19] | **Precisão** | **Cobertura** | ***F-measure*** |
|---:|---:|:---:|:---:|:---:|
| Cartão Cidadão (1) | 1 206 | 100,00% | 91,33% | 95,47% |
| Crime (2) | 953 | 77,45% | 33,76% | 47,02% |
| Email (1) | 712 | 98,73% | 89,66% | 93,98% |
| Estado Civil (2) | 56 | 71,43% | 58,82% | 64,52% |
| Filiação (2) | 184 | 58,06% | 31,03% | 40,45% |
| Habilitação Literária (2) | 738 | 71,11% | 14,35% | 23,88% |
| IBAN (1) | 82 | 100,00% | 93,33% | 96,55% |
| Localidade (2) | 1 060 | 47,55% | 22,15% | 30,22% |
| Morada (3) | 270 | 90,32% | 43,75% | 58,95% |
| NIB (1) | 235 | 100,00% | 92,00% | 95,83% |
| NIF (1) | 383 | 95,24% | 93,46% | 94,34% |
| NISS (1) | 778 | 100,00% | 90,56% | 95,04% |
| Nome (2) | 1 463 | 46,44% | 45,23% | 45,83% |
| Profissão (2) | 474 | 79,59% | 53,79% | 64,20% |
| Religião (2) | 25 | 72,73% | 66,67% | 69,57% |
| Saúde (2) | 452 | 67,14% | 34,56% | 45,63% |
| Telefone (1) | 506 | 96,61% | 82,61% | 89,06% |
| Telemóvel (1) | 601 | 97,18% | 90,20% | 93,56% |

Relativamente aos resultados, de entre os vários módulos criados para classificar os diferentes tipos de dados, apenas os módulos “Nome” e “Localidade” apresentam uma precisão inferior aos 50%, da

[18] A numeração à frente de cada tipo de dado representa o modelo usado na pesquisa: (1) expressões regulares; (2) dicionários *XML* e (3) padrões.

[19] Quantidade de dados pessoais conhecidos

análise observa-se que o baixo valor se deve maioritariamente a termos ambíguos entre os dois módulos, registando-se muitos nomes de pessoas com nomes de localidades e vice versa, assim como se verifica a existência de muitos registos duplicados promovidos pela utilização de contrações de preposição e artigos definidos nos termos compostos (que apresentam mais do que uma palavra) e a aplicação não conseguir lidar adequadamente com estas situações.

Quanto aos baixos índices de cobertura (inferiores a 50%), verifica-se a existência de um maior número de módulos. Em todos os sete casos, os baixos valores resultam da dificuldade da variabilidade linguística, concretamente na utilização de abreviaturas ou acrónimos para expressar as diferentes entidades. De entre os sete casos, a cobertura para reconhecimento de “habilitações literárias” e de “localidades” ficaram abaixo dos 25%, valores inferiores a um quarto. Em ambos os casos observa-se uma elevada disparidade de representação escrita, no caso das habilitações literárias na sua maioria não conferem o diploma que regula os vários níveis de habilitações literárias, já em relação às localidades para além da variabilidade linguística verifica-se igualmente muitas ambiguidades, com existência de muitas localidades a serem classificadas como nomes.

Comparando o gráfico da Figura 5.2 (que ilustra os resultados obtidos pelo *PerDa2Disco*, sendo o eixo das abcissas (x) a quantidade dos dados classificados e o eixo das ordenadas (y) os diferentes atributos) com a segunda coluna da Tabela 5.1, quantidade dos dados pessoais conhecidos, dá para perceber o efeito da precisão e da coberta dos modelos, principalmente através da análise da classificação de “Nomes” que apresenta mais dados do que existem na realidade existem, e pela classificação de “Localidades” (DesignacaoCP) apresentando menos dados do que existem.

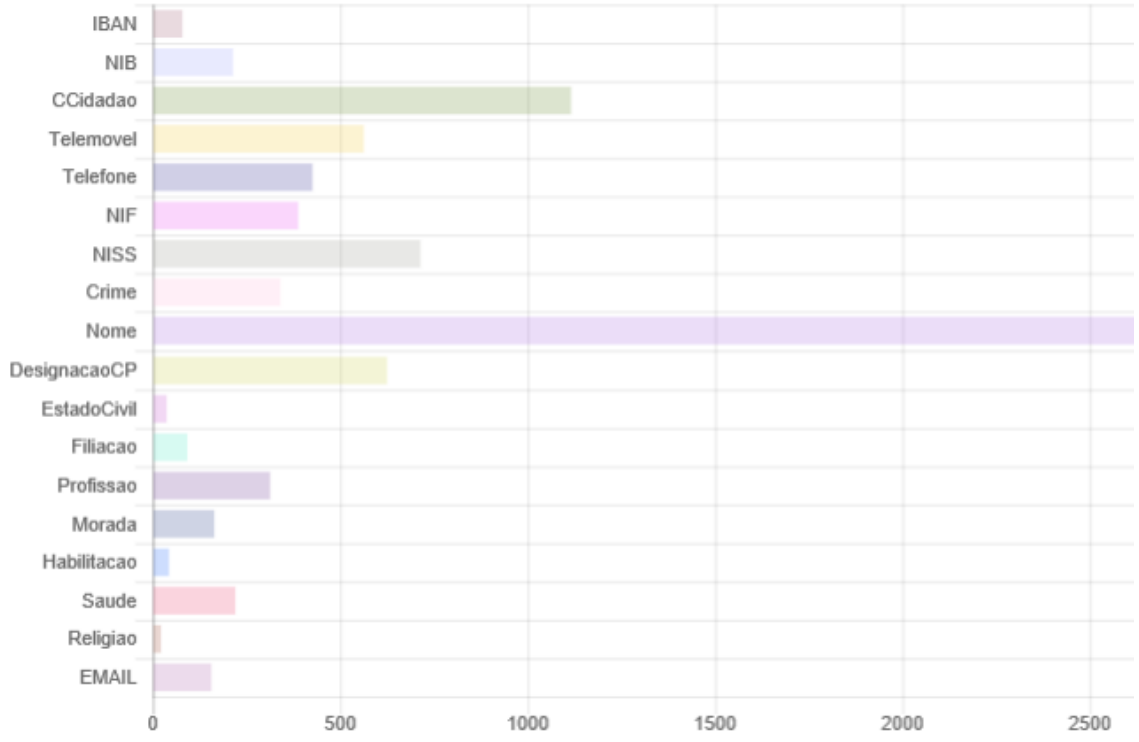



Figura 5.2: Demonstração gráfica dos resultados

Finalmente, os tipos de dados pessoais que são classificados com recurso às expressões regulares, embora no geral apresente bons resultados, parte do insucesso ao nível de cobertura deve-se à segregação na representação numérica dos números de telefone, telemóvel, NIF ou NIB.

Para a avaliação do desempenho, entre os diferentes modelos criados foram calculados os tempos de processamento sobre os mesmos 30 documentos utilizados na avaliação de qualidade, com a sua execução sendo sempre realizada no mesmo computador com as seguintes características:

- Processador: Intel® Core™ i5-6200U CPU @ 2.30GHz 2.40 GHz
- RAM: 16 GB

Tabela 5.2: Tempos de processamento (em segundos)

| Teste | Modelo utilizado | Configuração do Tipo de dados a procurar | Tempo (segundos) |
|---|---|---|---|
| 1 | *Regex* | IBAN, NIB, Cartão Cidadão, Telefone, Telemóvel, NIF, NISS e E-mail | 40,40 |
| 2 | Dicionário | Crime, Nome, Local, Estado Civil, Filiação, Profissão, Religião, Saúde e Habilitações | 37,30 |
| 3 | MaxEnt | Nome | 751,44 |
| 4 | Padrão | Crime, Nome, Local, Morada, Filiação, Profissão, Saúde e Habilitação | 31 479,04 |
| 5 | *Regex* + Dicionário | Combinação da configuração do teste 1 e do teste 2 | 66,76 |
| 6 | *Regex* + Dicionário + Padrão | Combinação da configuração do teste 1, 2 e apenas a Morada do teste 4 | 14 143,33 |

Fruto dos resultados obtidos, a configuração base utilizada para os estudos de caso foi de acordo com a Tabela 5.1, mas como iremos verificar foram igualmente realizados outros testes com diferentes configurações para medir o nível de desempenho em ambiente operacional aferindo os resultados de qualidade com recurso a uma pequena amostra aleatória.

# 5.2 Estudo de caso

Com intuito de testar a funcionalidade do artefacto em ambiente operacional foram realizados dois estudos de caso em duas organizações cuja área de negócio é bastante distinta, na Marinha Portuguesa e na empresa Link Consulting.

## 5.2.1 Marinha

Os testes realizados limitaram-se à secretaria partilhada da STI que serve o Gabinete da STI e a três direções que compõem o sector funcional das TI, a Direção de Tecnologias de Informação e Comunicações (DITIC), a Direção de Análise e Gestão da Informação (DAGI) e o Centro de Documentação, Informação e Arquivo Central da Marinha (CDIACM).

Os testes incidiram na componente dos dados não estruturados, em virtude de já estar prevista a realização de testes na componente estruturada na empresa Link Consulting e por existir um maior interesse na potencialidade da ferramenta nos dados não estruturados. Pretendeu-se avaliar de que modo a ferramenta poderia ajudar a identificar eventuais desvios de comportamento no tratamento de dados, assim como perceber em que medida existiam dados pessoais fora dos repositórios não estruturados (servidores de partilha de dados). Assim, na Marinha apenas se pretende testar a primeira componente do ciclo de tratamento de dados: a descoberta e classificação dos dados pessoais localizados. Foram estabelecidas três metas de avaliação:

(1) Verificar os tempos de execução dos diferentes modelos desenvolvidos em situações de operação remota e local;
(2) Registar a dimensão de dados pessoais descobertos por intermédio dos diferentes modelos e aferir a sua classificação, e por último;
(3) Perceber se existe algum desvio de comportamentos no tratamento de dados pessoais em relação à área de negócio da secretaria da STI.

Face ao exposto, no total foram realizados testes no repositório de dados do servidor da STI e a três áreas de trabalho dos computadores da secretaria com acesso ao servidor referido, incidindo apenas em documentos no formato das ferramentas de produtividade da *Microsoft* (nomeadamente *Word*, *Excel* e apresentações *PowerPoint*) e da *Adobe Systems* em particular o formato PDF em virtude de serem os tipos de documentos mais utilizados no funcionamento da secretaria. À semelhança dos testes em laboratório, a avaliação é centrada numa componente de desempenho e qualidade.

A amostra dos dados analisados no repositório de rede correspondeu aproximadamente a 4% do volume total, ou seja, 3 000 documentos em 69 280. A pesquisa foi efetuada através da configuração de rede e de modo normal de operação, ou seja, no decorrer das atividades diárias. Já análises efetuadas às estações de trabalho foram realizadas localmente por intermédio de uma *pen drive* e a amostra correspondeu igualmente a 3 000 documentos existentes na área de trabalho (12% do volume total), contudo registou-se que dois dos três computadores analisados apresentavam 944 e 1 557 documentos, tendo sido, nestes casos a amostra de 100%.

Para a **avaliação de desempenho** foram executados três testes (execuções) com os diferentes modos de descoberta disponíveis, com a particularidade da consulta apenas ter sido efetuada para descobrir os “nomes” e outros dados pessoais que fazem parte do modo *Regex*, tais como: cartão do cidadão, NIF, NISS, telemóvel, telefone, entre outros.

- Execução 1: Expressões regulares e dicionários (atributo “nome”);
- Execução 2: Expressões regulares e MaxEnt (atributo “nome”);
- Execução 3: Expressões regulares e padrões (atributo “nome”);

Foram separados em duas metas distintas para medir dois tempos de execução diferentes, a primeira (1) para perceber o tempo que a ferramenta demora a localizar todos os documentos de acordo com os formatos selecionados e a segunda (2) para conhecer o tempo de execução da análise na identificação e reconhecimento dos dados pessoais.

Com base nos testes executados, embora o volume de dados do repositório de rede apresente um volume significativamente maior do que os repositórios locais, constatou-se que para a primeira meta (1) os tempos de execução para a localização dos documentos é ligeiramente superior na pesquisa por rede, em relação a uma pesquisa local (como seria de esperar). No entanto, e embora o volume de documentos seja díspar entre os vários repositórios, pode-se aferir qual é a média por segundo na localização por documento (ver Tabela 5.3). O interessante deste teste prende-se com dar a conhecer o tempo aproximado que a ferramenta demora a fazer o levantamento de arquivo em relação ao tipo de acesso selecionado.

Tabela 5.3: Relação do tempos de localização dos documentos

| | **Rede** | **PC 1** | **PC 2** | **PC 3** | **Média PC** |
|---|---|---|---|---|---|
| Número total de documentos | 70 179 | 1 557 | 944 | 24 453 | 8 984,67 |
| Média localização (segundos) | 2 319,96 | 16,29 | 5,36 | 75,62 | 32,42 |
| Média por documento (segundos) | 0,0331 | 0,0105 | 0,0057 | 0,0031 | 0,0063 |
| (milisegundos) | 33,058 | 10,464 | 5,673 | 3,092 | 6,341 |

Já para a segunda meta (2), a relação dos tempos de execução da análise dos documentos verifica-se que a utilização de dicionários com expressões regulares apresenta claramente um melhor desempenho, comparativamente aos restantes modelos, seguindo-se a combinação do modo MaxEnt com as expressões regulares e por último a combinação da utilização de padrões com as expressões, 0,40 segundos por documento, 5,11 segundos por documento e 20,56 segundos por documento (ver Tabela 5.4), respetivamente.

Tabela 5.4: Relação do tempos de análise dos documentos por segundo

| | | **Execução 1 (*Regex* + Dicionário)** | **Execução 2 (*Regex* + MaxEnt)** | **Execução 3 (*Regex* + Padrão)** |
|---|---|---|---|---|
| **Repositório** | **Amostra (# Doc.)** | **Análise/Doc.** | **Análise/Doc.** | **Análise/Doc.** |
| Rede | 3 000 | 0,63 | 5,00 | 18,33 |
| PC 1 | 1 557 | 0,15 | 1,63 | 6,54 |
| PC 2 | 944 | 0,37 | 10,47 | 38,05 |
| PC 3 | 3 000 | 0,47 | 3,33 | 19,33 |
| Média PC | 1 834 | 0,33 | 5,14 | 21,31 |
| Média Geral | 2 125 | **0,40** | **5,11** | **20,56** |

Embora o modo de descoberta por dicionários e MaxEnt seja por intermédio da segmentação do texto por palavras, a combinação das expressões regulares com dicionários revelou-se ser a que apresenta um melhor desempenho. Considera-se que este facto se deve ao motivo pelo qual a análise em si é efetuada, ou seja, enquanto que o modo por dicionário utiliza a biblioteca *NameFinder* que se baseia puramente na identificação e reconhecimento dos termos existentes entre o texto a analisar e o dicionário, já no modo MaxEnt utiliza a biblioteca *NameFinderME,* significando que a análise que está a ser efetuada baseia-se numa distribuição probabilística, de acordo com o algoritmo do *Maximum*

*Entropy Model*. Finalmente para o modo padrão, o fraco desempenho, comparativamente com os outros modos, prende-se com a questão da sua análise ser efetuada caractere a caractere.

O objetivo da **avaliação de qualidade** consiste em perceber o comportamento da ferramenta num cenário de produção onde se desconhece a totalidade do tipo de dados existentes, para de acordo com os resultados avaliar, por amostragem, os registos *(logs)*, mesmo desconhecendo o conteúdo dos documentos procurar analisar o volume de dados duplicados e registos errados, para de certo modo aferir o nível de certeza entre o total de dados descobertos de um determinado tipo de dados e a precisão do módulo. No que diz respeito à identificação de palavras ambíguas, não foi possível analisar porque carecia de verificação do contexto do documento analisado.

Deste modo, a avaliação de qualidade foi faseada em duas componentes. A primeira (1) com recurso aos resultados obtidos nas três execuções realizadas, para aferir o nível de desempenho, analisando apenas os resultados do módulo “nome” dos vários modos de descoberta (conforme a seção 4.2.1, ver página 40); A segunda (2) através de uma quarta execução mediante a configuração de acordo com a Tabela 5.1 apresentada nos testes de laboratório (ver página 55) para permitir apresentar os resultados aos responsáveis do setor da Marinha.

Na primeira componente (1), verificou-se que o modelo padrão foi o modelo que em todos os testes realizados, quer em rede quer localmente nos computadores, identificou e reconheceu mais entidades, seguindo-se o dicionário e por último o MaxEnt, respetivamente, 70 538, 22 304 e 2 963, como se pode observar através do gráfico da Figura 5.3 . No entanto, identificar mais entidades não significa que seja o melhor, pois no volume total das entidades reconhecidas poderão existir registos duplicados, ou seja, identifica termos compostos de entidades que poderá estar a referir dois ou mais termos como pertencentes a uma única entidade. Não obstante, como primeira abordagem dará para ter uma ideia sobre a existência de dados pessoais e o volume que representam em relação aos restantes dados.

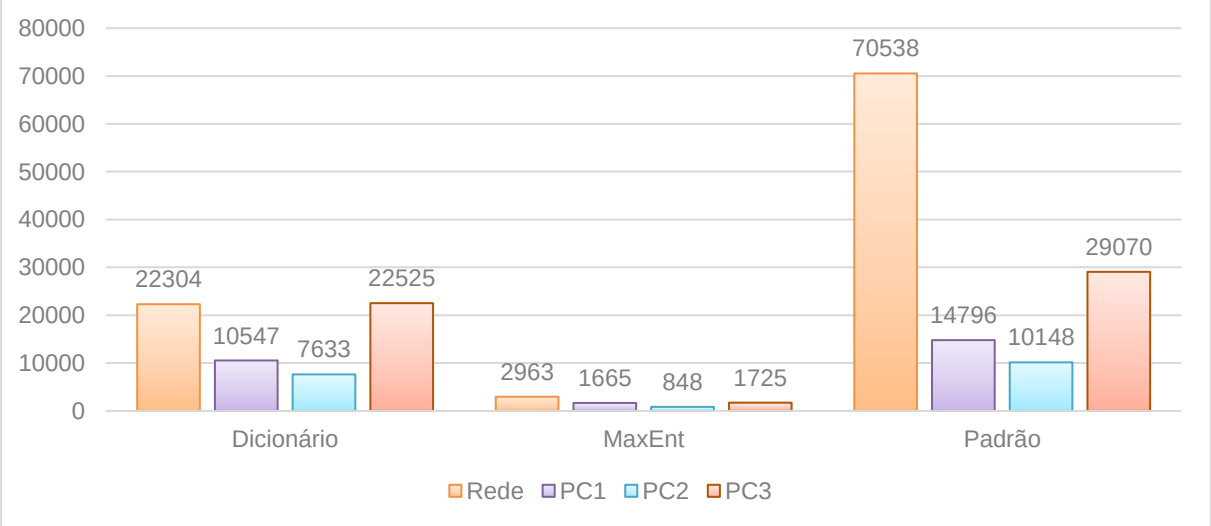



Figura 5.3: Nomes identificados

Embora a Figura 5.3 represente os nomes identificados de todos os repositórios, apenas foram analisados os dados referente ao servidor de rede. O modelo padrão identificou 70 538 registos como

sendo “nomes”, confirmando-se no estudo efetuado, a existência de registos duplicados principalmente nas situações onde o nome contém contrações da preposição (“da”, “de” e “do”). Neste sentido, o conjunto de dados foi trabalhado para se poder identificar registos duplicados. Verificou-se como possíveis registos duplicados um total de 4 522 registos, distribuídos da seguinte forma: (“de”) 3 184; (“da”) 1 209; e (“do”) 129, podendo existir mais registos duplicado, em virtude de apenas se ter analisado uma das situações em que pode ocorrer duplicação de termos. recomendando-se a revisão da limitação apresentada na seção 4.2.6, Consulta (ver página 45).

Já em relação ao modelo dicionário que apresenta como resultado 22 304 registos, conclui-se que o modelo apresenta como registos duplos, sensivelmente, 1 889 registos que dizem respeito aos termos que possuam entre si uma contração da preposição, conjunção coordenativa ou artigo definido. Para o despiste da análise foram avaliados todos os termos (1) com apenas um elemento que estão separados do próximo termo com uma distância de uma palavra, (2) com dois elementos cuja separação são duas palavras do terceiro termo e (3) com quatro elementos que distam de três palavras do termo seguinte, registando-se respetivamente, de 993, 641 e 255.

Finalmente, o modo MaxEnt reconheceu um total de 2 963 entidades no repositório de rede, verificando-se 76 registos não válidos. É referido que são registos não válidos, em virtude do aspeto dos nomes detetados que são compostos por um aglomerado de caracteres sem sentido (“IIgl&fl'”, “Edlllciolln'”, “QWmtl'di"doil”, “QiUddidili“, “P~gl”, …). Em relação aos registos duplicados, foi efetuada uma análise análoga ao modelo dicionário, registando-se apenas 66 possíveis registos.

Para a segunda componente (2) da avaliação de qualidade, como referido foi mediante um quarto teste (execução) de acordo com a configuração dos resultados de precisão e cobertura apresentados nos teste de laboratório (ver Tabela 5.1, página 55), onde se efetua uma análise comparativa por amostragem, relacionando os resultados obtidos com a precisão do respetivo módulo, finalizando com uma avaliação qualitativa por intermédio de entrevistas e questionários realizados aos responsáveis pelo setor da Marinha, após a realização de uma demonstração prática da ferramenta e de uma apresentação do resultados obtidos.

Infelizmente, não foi possível aferir a qualidade do módulo “data” e do “código postal” das expressões regulares e como estes termos não têm nenhuma função associada de validação, motivo pelo qual foi efetuada uma análise aos resultados *(logs)* para apreciar a veracidade deste tipo de dados. No repositório de rede, dos 20 443 registos identificados como sendo “datas”, poder-se-á afirmar pelo tipo de resultado que (1) existe uma probabilidade muito elevada de 17 837 serem efetivamente datas, porque de acordo a análise manual aos registos *(dataset* dos *logs)* foi possível verificar que apresentavam um formato que se assemelhava a datas (e.g. 27/12/1982); (2) 2 353 claramente que não o são (falsos positivos), pois apresentavam um formato diferente do que seria expectável (sequência de apenas cinco algarismos) e (3) 253 poderão suscitar dúvidas devido a terem um formato igualmente sequencial, mas composto por oito algarismos 8 (e.g. 20180429).

Na sequência desta primeira analise, foi solicitado acesso a 10 documentos que apresentam dados do tipo “data” para fazer uma comparação manual, concluindo-se que todos os registos do formato em (1)

diziam de facto respeito a datas, verificou-se que constavam no registo de datas; os registos que se apontam como falsos positivos (2) correspondem a números mecanográficos de militares e no caso (3) correspondiam a uma parte do nome do assunto de análises técnicas com o objetivo de indicar a data da consulta pública. Verificou-se ainda dois documentos, que para além das datas localizadas continham a data escrita por extenso, mas como referido na seção 4.2.4 o módulo não está habilitado a reconhecer este tipo de formato. Já para o tipo de dado "código postal", segundo os registos, todas as entidades descobertas estavam de acordo com o padrão e da análise manual efetuada, igualmente, a 10 documentos, diferentes dos anteriores, constatando-se uma correspondência de 100%.

Outro tipo de dado que não foi aferido a qualidade é o "género", mas face aos poucos resultados obtidos foi possível comparar manualmente todos os documentos em que este tipo de dado foi localizado, confirmando-se que todos os registos detetados correspondem efetivamente a situações que pretendem identificar o "género", mas a realidade é que se desconhece o nível de precisão e cobertura, pelo que poderá existir muito mais situações que não foram detetadas.

Tabela 5.5: Resultados do estudo de caso Marinha

| Tipo de Dado | P (%) | C (%) | Rede | | PC1 | | PC2 | | PC3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | D | PD | D | PD | D | PD | D | PD |
| Cartão Cidadão | 100,00 | 91,33 | 0 | 0 | 1 | 1 | 4 | 4 | 0 | 0 |
| Código Postal | N/V | N/V | 1 405 | 1 405 | 400 | 400 | 340 | 340 | 493 | 493 |
| Crime | 77,45 | 33,76 | 16 | 37 | 7 | 16 | 104 | 239 | 95 | 218 |
| Data | N/V | N/V | 20 443 | 17 837 | 5 466 | 4 445 | 2 118 | 1 887 | 22 255 | 20 068 |
| Localidade CP | 47,55 | 22,15 | 12 760 | 27 392 | 2 413 | 5 180 | 1 405 | 3 016 | 6 114 | 13 125 |
| EMAIL | 98,73 | 89,66 | 379 | 417 | 216 | 238 | 35 | 39 | 180 | 198 |
| Estado Civil | 71,43 | 58,82 | 13 | 16 | 3 | 4 | 12 | 15 | 55 | 67 |
| Filiação | 58,06 | 31,03 | 418 | 782 | 161 | 301 | 347 | 649 | 936 | 1 751 |
| Género | N/V | N/V | 24 | N/V | 10 | N/V | 14 | N/V | 29 | N/V |
| Habilitação | 71,11 | 14,35 | 341 | 1 690 | 164 | 813 | 131 | 649 | 466 | 2 309 |
| IBAN | 100,00 | 93,33 | 4 | 4 | 0 | 0 | 0 | | 0 | 0 |
| Morada | 90,32 | 43,75 | 1 387 | 2 863 | 423 | 873 | 367 | 758 | 771 | 1 592 |
| NIB | 100,00 | 92,00 | 9 | 10 | 0 | 0 | 0 | | 6 | 7 |
| NIF | 95,24 | 93,46 | 118 | 120 | 43 | 44 | 17 | 17 | 103 | 105 |
| NISS | 100,00 | 90,56 | 10 | 11 | 0 | 0 | 4 | | 1 | 1 |
| Nome | 46,44 | 45,23 | 22 304 | 22 901 | 10 547 | 10 829 | 7 633 | 7 837 | 22 525 | 23 128 |
| Profissão | 79,79 | 53,79 | 2 821 | 4 185 | 1 760 | 2 611 | 1 608 | 2 385 | 4 579 | 6 792 |
| Saúde | 67,14 | 34,56 | 2 | 4 | 7 | 14 | 4 | 8 | 10 | 19 |
| Telefone | 96,61 | 82,61 | 639 | 747 | 76 | 89 | 87 | 102 | 923 | 1 079 |
| Telemóvel | 97,18 | 90,20 | 458 | 493 | 24 | 26 | 241 | 260 | 237 | 255 |

A Tabela 5.5, acima, demonstra o volume dos tipos de dados descobertos nos diferentes repositórios, indicando a relação do volume dos possíveis dados, cujo cálculo foi baseado na precisão e cobertura dos respetivos módulos, de acordo com a seguinte formula:

$$PossiveisDados\ (PD) = \frac{DadoDescoberto\ (D) * Precisão\ (P)}{Cobertura\ (C)} \quad (9)$$

De acordo com os mini-questionários efetuados (ver Tabela 5.6, abaixo) aos principais intervenientes da STI e igualmente responsáveis por garantir a aplicação das medidas tecnológicas e de segurança na Marinha, complementados com as entrevistas realizadas (ver Apêndice E, Resumo de entrevistas), pode-se concluir que no geral a avaliação em relação ao protótipo desenvolvido, no geral é bom, sendo unânime que tratando-se de um trabalho de índole académico, mas com uma forte componente aplicacional, este superou as expetativas iniciais. Embora se verifique que não esteja perfeito e que requeira de alguns melhoramentos, especialmente em alguns módulos para descoberta de dados pessoais, o principio está bem patente e poderá ser claramente uma ajuda para qualquer organização, para contribuindo para a segurança e privacidade dos dados, seja para uma organização privada ou da administração pública.

Tabela 5.6: Resultado do mini-questionário

| **Questão**[20] | **Questionário** | | | | | | **Média** |
|---|---|---|---|---|---|---|---|
| | **# 1** | **# 2** | **# 3** | **# 4** | **# 5** | **# 6** | |
| **1** | 5 | 5 | 3 | 4 | 3 | 2 | 3,67 |
| **2** | N/A | 4 | 4 | 5 | N/A | N/A | 4,33 |
| **3** | 4 | 5 | 4 | 4 | 4 | 3 | 4,00 |
| **4** | 5 | 4 | 5 | 4 | 5 | 5 | 4,67 |
| **5** | 5 | 5 | 5 | 4 | 4 | 4 | 4,50 |
| **6** | Sim | Sim | Sim | Sim | Sim | Sim | 5,00 |
| **7** | 5 | 5 | 5 | 4 | 4 | 4 | 4,50 |
| **Média** | 4,65 | 4,55 | 4,23 | 4,06 | 3,88 | 3,50 | 4,15 |

A classificação atribuída dos dados pessoais e a forma de como é atribuído o seu grau, foi considerado um ponto forte, pois simplifica o processo de definir prioridades, alertando para os dados que poderão ser mais problemáticos.

Em relação aos resultados obtidos da análise efetuada aos repositórios da STI, no que concerne ao tipo de dados pessoais descobertos está de acordo com o inicialmente previsto, pois verifica-se que os tipos de dados descobertos correspondem ao tipo de dados do normal funcionamento da secretaria do setor das TI da Marinha [79]. Embora seja requerida uma análise mais profunda dos registos comparando-os com os documentos em si, mediante os resultados apresentados a proporção do volume de dados em relação ao tipo de dados descobertos fazem sentido. Claramente que se estaria à espera de encontrar maioritariamente nomes, locais e datas pelo motivo que os documentos que a secretaria gere no seu dia a dia são tipicamente documentos relacionados com circulares internas, deslocações em serviço, propostas de aquisição de material e todos estes documentos contêm obrigatoriamente nomes, datas e localidades.

[20] Ver detalhe no Apêndice D, Guião de entrevista e questionário

### 5.2.2 Link Consulting

Outro estudo de caso, com testes operacionais que foram realizados para aferir o comportamento da ferramenta num cenário real, foi realizado na empresa Link Consulting através da avaliação do sistema *Edoclink*. Este sistema consiste numa solução integrada de Gestão Documental e de suporte a processos de decisão, desenvolvida para ambientes *web*, que assegura as atividades de registo e encaminhamento de documentos num contexto de gestão de processos internos, podendo recorrer a formulários eletrónicos associados a encaminhamentos predefinidos.

Os testes formam realizados em duas fases distintas: uma primeira (1) em ambiente de qualidade onde foram dadas permissões de acesso para explorar a estrutura e conectividade à base de dados e a segunda (2) em ambiente de produção, cujos testes foram realizados diretamente por intermédio dos responsáveis pela área.

Numa fase preparatória, foi identificada a necessidade de efetuar alguns ajustes (melhoramentos) de forma a criar uma agregação dos resultados. Em vez de apresentar o tipo de formato do documento deveria ser apresentado a área de negócio, pelo que foi necessário perceber conceptualmente o sistema e como é que estava estruturado. Verificou-se que os documentos estão organizados por registos e/ou processos pertencentes a uma determinada pasta documental (ver Figura 5.4), cujos conceitos base são [80]:

- Pasta documental: A tradicional arrumação dos documentos foi reutilizada pelo mundo digital, com as pastas que conhecemos de todos os sistemas operativos. Quando falamos de documentos na ótica da sua gestão, as pastas continuam a ser a forma de os "arrumar". No entanto no *Edoclink*, estas pastas podem ser muito mais ricas, pois permitem guardar não só os documentos, mas também os fluxos de encaminhamento dos documentos, assim como os mais variados anexos, gerados pelos utilizadores ou automaticamente por aplicações. As pastas documentais são agregadas por classe, que da mesma forma que o livro, tipificam conjuntos de pastas processuais.
- Fluxo processual: representação do fluxo de um documento na organização, sendo que a vida de um documento corresponde a sucessivas etapas, algumas executadas por humanos, outras realizadas automaticamente por aplicações diversas, ou mesmo executadas por entidades externas como parceiros e clientes. As etapas podem ser predefinidas, por exemplo por razões normativas, ou serem totalmente decididas por um utilizador que define a etapa seguinte mais adequada. Em alguns casos, os fluxos são condicionados por regras baseadas em dados do próprio registo. Um conjunto de etapas predefinidas formam o conceito de fluxo processual predefinido.
- Documento: qualquer agregação de informação com valor documental que pode ter qualquer formato eletrónico, tamanho ou formatação e pode mesmo ser um documento tradicional em papel que o sistema passa a referenciar. Um documento passa a ser tratado pelo *Edoclink* quando é registado. O registo não é mais que criar um descritivo (meta-data) que captura os aspetos mais relevantes para a gestão do documento. Na prática, os documentos são os seus

registos, sendo a informação constante do documento guardado num ficheiro que o registo indexa (normalmente designado por livro).


Figura 5.4: Conceito geral do sistema *Edoclink* [80]

Ao nível físico, pode considerar-se que o *Edoclink* tem dois repositórios distintos: um de dados estruturados onde são persistidos meta-dados e outro não estruturado onde são persistidos documentos (ficheiros). Os documentos propriamente ditos estão descontextualizados e não têm significado semântico sem o enquadramento dos meta-dados. Desta descontextualização surge a necessidade de apresentar os resultados da informação recolhida pela execução da aplicação *PerDa2Disco* sobre repositórios de documentos (dados não estruturados) com enquadramento dos conceitos representados, anteriormente, no *Edoclink*.

Com vista a enquadrar funcionalmente os documentos encontrados nos repositórios não estruturados do *Edoclink*, foi desenvolvido um módulo de contextualização da informação recolhida. Desta forma, torna-se possível incrementar a funcionalidade do *PerDa2Disco* sem alterações, quer na própria ferramenta quer nos sistemas em análise, e capacitar os resultados obtidos de contexto funcional para melhor entendimento dos processos e áreas de negócio dos quais resultam os documentos encontrados. Os principais requisitos deste módulo são:

- Utilizar a mesma configuração que o *PerDa2Disco*, para evitar a duplicação de configuração;
- Devolver como resultado um ficheiro em formato .csv, para permitir a importação do resultado final para o EAPY;
- Não requerer alterações à ferramenta *PerDa2Disco*, para evitar dependências e facilitar evoluções.

O módulo obtém a identificação do documento e com esta identificação efetua três pesquisas para obtenção do contexto funcional do documento:

- Classe de pasta documental: classes das pastas documentais nas quais o documento está inserido. Esta relação tem dois níveis de indexação pois as pastas documentais não contêm documentos, apenas registos e fluxos processuais;
- Fluxo processual predefinido: conjunto de etapas nas quais o documento está inserido.
- Livro: é relevante salientar que os documentos podem ter múltiplas associações, ou seja, o mesmo documento pode estar inserido em vários registos ou distribuições.

O método criado, *PerDaEdoclink*, utiliza como *input* o resultado da ferramenta *PerDa2Disco*, no formato .csv, e acrescenta o contexto funcional conforme descrito acima, tornando o resultado final compatível com a importação para a aplicação de governação EAPY. Isto foi possível porque o sistema do *Edoclink* para além de estar organizado por processos de negócio o nome do ficheiro está padronizado permitindo conhecer bem quais os carateres responsáveis por identificar a área de negócio (pasta documental, registos e/ou fluxos processuais).

Como referido, na primeira fase, todos os testes foram realizados no ambiente de qualidade permitindo explorar a conetividade a uma base de dados do tipo *SQLServer* e conhecer a estrutura do repositório de dados não estruturados. Esta fase revelou-se bastante pertinente processo porque permitiu detetar e corrigir algumas lacunas no processo de análise dos documentos, nomeadamente a situação de alteração do estado de um determinado documento no decorrer da execução da análise e/ou quando está a escrever para os ficheiros de relatório (.csv), algo que a aplicação original não previa e nos testes de laboratório e da Marinha não se registou esta coincidência. Isto é, no caso de (1) um utilizador abrir um documento durante o processo de análise desse mesmo documento por parte do protótipo, a ferramenta interrompe a análise do documento mas dá continuidade para a análise de outro documento não bloqueando a execução; assim como se (2) a análise já tiver sido concluída mas no momento de escrever para os ficheiros de registo (.csv) ocorrer a alteração do estado (e.g. o utilizador entretanto abriu o documento mantendo-o aberto), o protótipo ignora o registo desse documento continuando para o seguinte mas a sua execução não é bloqueada.

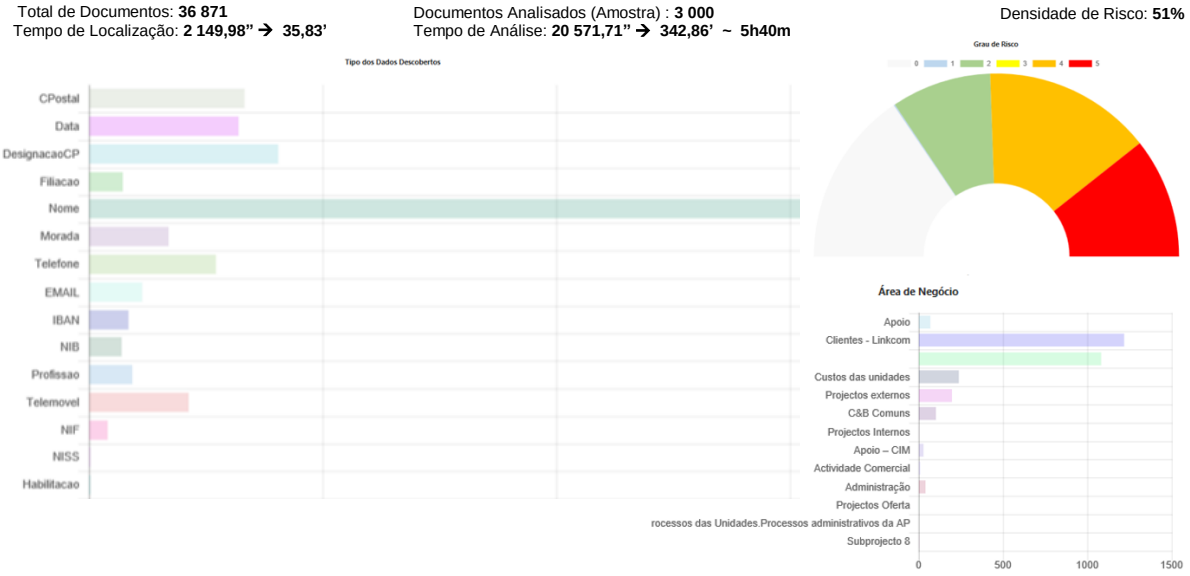
Para a segunda fase, o protótipo foi preparado e entregue aos responsáveis pelo ambiente de produção do *Edoclink*, para executarem a ferramenta num ambiente não estruturado e estruturado. Neste estudo de caso pretendeu-se apenas identificar e classificar os vários tipos de dados, ficando a análise de qualidade a ser realizada posteriormente pelos responsáveis do sistema. Em relação aos **dados não estruturados**, a grande diferença da apresentação dos resultados, como anteriormente referido, consiste numa agregação por processo de negócio. Os testes foram realizados mediante a mesma configuração do teste realizado na Marinha e corresponde à apresentada na Tabela 5.1 (ver página 55), desconhecendo-se as características do computador e da latência da rede (tempos de resposta).

A Figura 5.5, abaixo, representa os resultados suspeitos na execução do protótipo no sistema *Edoclink*, na componente não estruturada. É possível verificar o número de dados analisados em relação ao volume total de documentos existentes no repositório, no caso, uma amostra aleatória de, sensivelmente, 8,14%. O tempo que demorou a localizar o volume total de documentos (cerca de 36 minutos) e tempo de análise da amostra (5 horas e 40 minutos), ou seja, a execução do teste demorou

no total 6 horas e 16 minutos, verificando-se uma densidade de risco a rondar os 51% cuja distribuição pelos documentos foi:

- 929 | Ausência de dados pessoais;
- 531 | Combinação de dados pessoais de categoria 1 (nomes, moradas, …);
- 895 | Combinação de dados pessoais de nível 1 com o nível 3;
- 638 | Documentos com suspeita de dados considerados especiais.

Nos gráficos de barras é possível ter uma perceção geral do tipo de dados descobertos e de qual é a área de negócio quais as áreas de negócio identificadas. E na tabela, é possível perceber quais são as áreas de negócio associadas a um determinado documento, quais são os tipos de dados pessoais existentes, a classificação de risco e a respetiva densidade de risco em proporção ao global.




| ◆rea Neg◆cio | Pasta | Ficheiro | Tipo de Dado | Clas. Risco | Densidade Risco |
|---|---|---|---|---|---|
| Apoio,ENT-LINK | \\psedoclink01.aitec.pt\D$\edoclink\instances\edoc\Store01\2012\10\23\d44f370c-73e5-4334-bb92-b520bc1c41f5\ | 922c5f4e-6145-4387-8798-33fc8fc31705.pdf | | 0 | 0.00000 |
| Clientes - Linkcom | \\psedoclink01.aitec.pt\D$\edoclink\instances\edoc\Store01\2015\2\19\7abb4057-93ce-40f4-87f8-2492b2d31d14\ | 905da5dc-c33b-469c-abbd-e5cd9c340e19.1..pdf | Filiacao, Morada, CPostal, Data, DesignacaoCP, Nome | 5 | 0.00033 |
| Clientes - Linkcom | \\psedoclink01.aitec.pt\D$\edoclink\instances\edoc\Store01\2015\1\27\ec4eb35b-a181-4de2-91c7-f68f5259af10\ | bf6d4089-b15d-4cef-b526-7c7f68cdcd32.1..pdf | Filiacao, Telefone, Morada, CPostal, Data, EMAIL | 5 | 0.00033 |

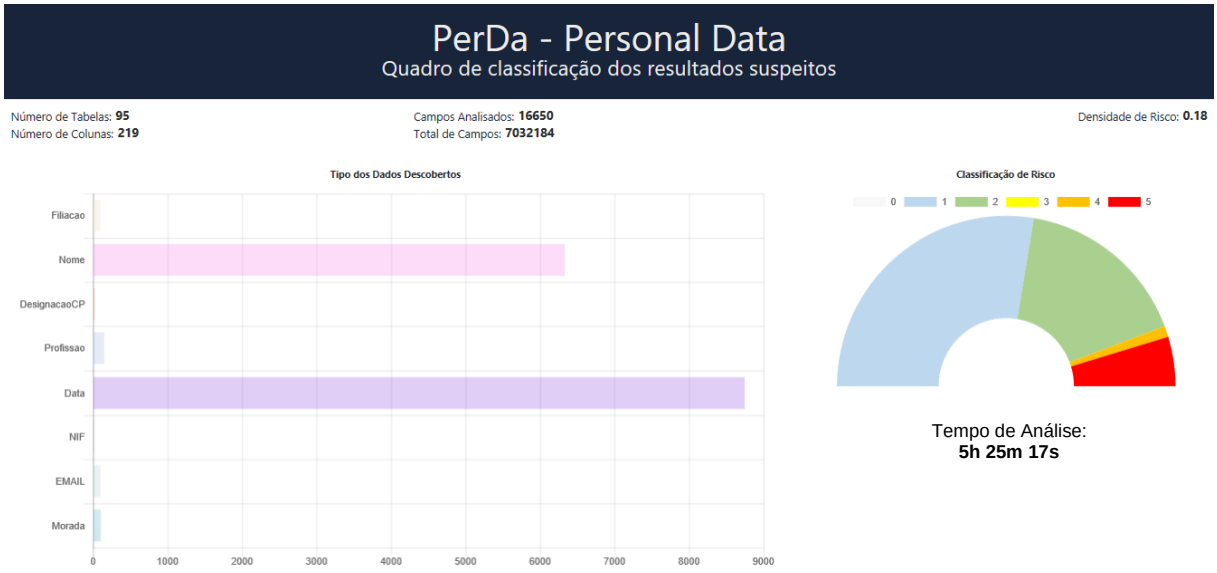
Figura 5.5: Resultados do teste *Edoclink* (não estruturado)

A Figura 5.6, abaixo, representa os resultados suspeitos da execução do protótipo ao sistema *Edoclink*, na componente estruturada. Mesmo não conhecendo a estrutura e dependências das tabelas é possível conhecer o número de tabelas e colunas existentes no total, assim como o número de campos (células) analisadas em relação ao número total de campos da base de dados, que no caso, representa uma ínfima parte, uma amostra com apenas 0,24%. Embora se calcule o tempo de execução, ao

contrário dos dados não estruturados este não é apresentado de uma forma automática. A densidade de risco, foi de 18%, com distribuição dos dados pessoais por tabela:

- 53 | Dados pessoais soltos de categoria 1 (nomes, moradas, …);
- 32 | Combinação de dados pessoais de categoria 1 (nomes, moradas, …);
- 02 | Combinação de dados pessoais de nível 1 com o nível 3;
- 09 | Documentos com suspeita de dados considerados especiais.

À semelhança dos resultados para os dados não estruturados, no gráfico de barras é possível ter uma perceção geral do tipo de dados descobertos e na tabela (em forma de lista), é possível conhecer o nome da base de dados (URL), perceber quais são os tipos de dados pessoais existentes, a classificação de risco e a respetiva densidade de risco em proporção ao global de cada tabela.




| URL | Nome Tabela | Tipo de Dados | Cals. Risco | Densidade Risco | Campos Analisados | Total Campos |
|---|---|---|---|---|---|---|
| sqlserver //PSVMSQL16.aitec.pt database=edoc | ACCESS_INHERITED_FROM_CODES | Filiacao, Nome | 5 | 0,000520833 | 44 | 44 |
| sqlserver //PSVMSQL16.aitec.pt database=edoc | ACL_PENDING_OPERATIONS | Filiacao, Data | 5 | 0,009895833 | 300 | 37428 |
| sqlserver //PSVMSQL16.aitec.pt database=edoc | ACTIONS | Filiacao | 5 | 0,000520833 | 10 | 10 |
| sqlserver //PSVMSQL16.aitec.pt database=edoc | ADMIN_LOGS | NIF, Data | 4 | 0,007916667 | 400 | 416124 |
| sqlserver //PSVMSQL16.aitec.pt database=edoc | ALARMS_ACTIVE_EVENT | DesignacaoCP, Nome, Email, Data | 2 | 0,003958333 | 500 | 1140 |

Figura 5.6: Resultados do teste *Edoclink* (estruturado)

Mesmo desconhecendo o conteúdo dos documentos (não estruturados) e/ou os registos (estruturados), através da análise dos registos *(logs)* foi possível verificar alguns falsos positivos, nomeadamente em

relação ao atributo “filiação”, e à semelhança do estudo de caso da Marinha, verifica-se alguns registos ambíguos entre os “Nomes” e as “Localidades”.

Na sessão realizada para a apresentação dos resultados aos responsáveis da Link Consulting [81], verificou-se que (1) para a componente dos dados não estruturados, estavam à espera de encontrar um equilíbrio entre os dados relacionados com nomes e datas, isto porque praticamente todos os documentos carregues no sistema são datados e assinados. Eventualmente, a quantidade de localidades (DesignacaoCP) poderia ser maior, mas pelo que foi apresentado parte dos atributos identificados como “nomes” podem ser localidades. Já em relação ao atributo “filiação”, para além dos falsos positivos referidos, parece que parte dos resultados estão a ser confundidos com organizações, pelo que o módulo poderá requerer uma afinação dos termos a identificar ou, eventualmente, alterar o nome do tipo de dado para “organização”; (2) para a componente de dados estruturados, embora a amostra dos campos analisados seja reduzida, tendo em conta que os tipos de dados que se está a procurar são respeitantes a pessoas particulares, os resultados parecem aceitáveis, contudo obtiveram-se muito poucas localidades.

# 6 Conclusão e trabalho futuro

A privacidade é um direito fundamental reconhecido na Declaração Universal dos Direitos Humanos. A evolução tecnológica conduziu a uma massificação dos dados, fazendo com que nas últimas décadas tenham surgido inúmeras normas e recomendações internacionais sobre segurança tecnologia e privacidade dos dados. Contudo, no que diz respeito à aplicabilidade da privacidade dos dados verifica-se a existência de múltiplas interpretações ao nível mundial, inclusive, pelos diferentes Estados-Membros da EU. O novo RGPD visa harmonizar a proteção dos direitos e liberdades dentro do espaço europeu, impondo regras muito claras a todos os organismos que tratam dados pessoais.

A presente dissertação de mestrado teve como objetivo estabelecer um referencial que contribua para a descoberta de dados pessoais sensíveis, que possam violar a privacidade dos indivíduos com suporte de um artefacto. A instanciação do artefacto foi efetuada através do desenvolvimento e implementação do protótipo de uma aplicação, e complementado com a definição de um método com os procedimentos com adequados de tratamento dos dados, de forma a contribuir para o cumprimento das disposições legais estabelecidas no RGPD.

Pretendeu-se com a dissertação explorar as técnicas de IE e NLP com um protótipo capaz de detetar, identificar e classificar os diferentes tipos de dados pessoais, alertando os responsáveis pelo tratamento de dados a tomarem consciência do volume de dados pessoais que os diferentes repositórios organizacionais (estruturados e não estruturados) possam eventualmente ter, conduzindo desta forma a uma reflexão sobre os riscos existentes no processo de tratamento de dados e de como é que poderão mitigar os riscos de uma possível violação da privacidade de dados pessoais.

O modelo seguido para promover a classificação NER foi o *Maximum Entropy*, recorrendo às ferramentas, utilitários e bibliotecas do *Apache OpenNLP* para a linguagem *Java*, promovendo a criação de novos modelos para o NER na língua portuguesa, com o intuito de descobrir dados pessoais de acordo com os atributos do RGPD. Sendo que, com o desenvolvimento da instanciação do protótipo tornou-se evidente que o desafio não passa apenas pela aplicação dos modelos existentes para o efeito, mas sim da forma como se enfrenta os constrangimentos da variabilidade linguística e ambiguidades da língua para se conseguir assegurar a veracidade e qualidade dos resultados, ou seja, os dados descobertos representam a mesma realidade dos dados originais, minimizando os falsos positivos e os falsos negativos.

Face ao problema identificado e tendo presente o objetivo principal de garantir a privacidade de dados, para avaliação dos resultados foi seguido um percurso capaz de responder ao segundo e terceiro objetivo específico, nomeadamente: (1) desenvolver um software que possa contribuir para a descoberta de dados pessoais, potencialmente sensíveis, suscetíveis de violar a privacidade de indivíduos, bem como (2) identificar uma metodologia para o tratamento de dados, de forma a garantir o cumprimento das disposições legais estabelecidas pelo RGPD.

Neste sentido, considera-se que os resultados obtidos foram bastantes satisfatórios, pese embora não tenha sido possível fazer uma comparação com outra aplicação ferramenta similar, isto porque dentro da disponibilidade não foram encontradas ferramentas de fonte aberta de características similares para a língua portuguesa, e as poucas disponíveis consistiam em demonstrações não sendo possível utilizar para uma análise ao nível de repositórios de dados (estruturados ou não-estruturados).

Por um lado, através dos testes em laboratório, foi possível aferir o rigor dos módulos NER, mediante uma base referencial, permitindo encontrar um bom equilíbrio da utilização dos vários modos de descoberta face à qualidade *versus* desempenho e conseguindo resultados acima dos 90% em alguns módulos NER. Estes resultados deveram-se em parte às melhorias introduzidas, quer pela forma modular como se conseguem realizar as consultas, quer pelos módulos NER criados para a localização de novos tipo de dados pessoais.

Para além da validação teórica dos módulos NER (precisão e cobertura) foi possível fazer uma validação em contextos operacionais (estudos de caso) da validade do protótipo. Nos testes foi possível corrigir situações, que de outra forma não teria sido possível, e verificar pelas várias entrevistas realizadas, que o protótipo desenvolvido tem utilidade na identificação e classificação de dados pessoais, sendo capaz de, numa primeira fase, dar a conhecer aos responsáveis pela privacidade de dados o ponto de situação sobre o tipo de dados existentes e em que locais da organização.

De acordo com os resultados apresentados, no estudo de caso da Marinha, o responsável máximo pelo funcionamento da secretaria [79], revê o tipo de dados descobertos como sendo os dados expectáveis. Claramente que se estaria há espera de encontrar maioritariamente nomes, locais e datas pelo motivo que os documentos que a secretaria gere no seu dia a dia são tipicamente documentos relacionados com circulares internas, deslocações em serviço, propostas de aquisição de material e todos estes documentos contem obrigatoriamente nomes, datas e localidades. Embora possam existir módulos com um nível de precisão inferior, globalmente considerando, para se ter uma fotografia do que existe em termos de dados pessoais é um bom indicador. Permite acelerar o processo porque conhecendo o rigor do módulo NER, verificando o nível de classificação de risco do dado e conhecendo a localização no repositório permite priorizar a investigação e ir diretamente aos locais de maior interesse.

Finalmente, através do estudo de caso do *Edoclink* da empresa Link Consulting, foi possível apurar o funcionamento do protótipo através da descoberta de tipos de dados em dados não estruturados e estruturados, concluindo que o protótipo embora careça de alguns melhoramentos, nomeadamente na questão da afinação dos falsos positivos, já apresenta potencialidades práticas. Contudo, tendo em conta a modularidade da configuração dos módulos relativos aos tipos de dados (independente da ferramenta) é possível reduzir os falsos positivos desde que os módulos sejam refinados. Foi verificado que consideram [81] como um aspeto forte do trabalho (1) a introdução do conceito de Densidade de Risco e (2) a forma como os resultados são apresentados, em virtude agregar toda a informação relevante para a execução de ações de proteção dos dados, permitindo fazer uma priorização dos aspetos mais importantes, facilitando o trabalho de diagnóstico.

Como trabalho futuro foi identificado a necessidade:

(1) de melhorar o processo de aprendizagem dos modelos em português. Embora tenham sido desenvolvidos esforços para incorporar no protótipo o processo de aprendizagem, não foi possível conseguir implementar esta funcionalidade. A ideia subjacente consiste em etiquetar os textos simultaneamente ao processo de leitura para verificar se existem dados pessoais onde o documento analisado serviria para alimentar os ficheiros binários (.bin) do modo MaxEnt criados onde seriam efetuados incrementos sucessivos à medida que se fossem realizando análises e validando a classificação da marcação dos dados pessoais mediante os resultados encontrados.

(2) de explorar a integração do modo de descoberta por dicionário e por padrão de modo a permitir melhorar a qualidade da precisão e cobertura, nomeadamente em relação aos termos duplicados, procurando igualmente desenvolver padrões para otimizar o nível de desempenho.

(3) de minimizar o número de ambiguidades poder-se-á criar um campo antes de executar a análise do repositório para se identificar como se pretende reconhecer determinados termos, possivelmente de acordo com a natureza do documento. Ou seja, se a análise incidir num repositório que a maioria dos documentos são *curriculum vitae*, muito provavelmente iremos classificar termos como “flores”, “branco”, “liberdade”, “céu”, entre outros como classificação “nome”.

(4) outro aspeto considerado muito importante, será tornar a ferramenta capaz de lidar com contrações de preposição (“do”, “de”, “da”, …), conjunção coordenativa (“e”) ou artigos definidos do tipo (“a”, “o”). Neste domínio concreto, poderá ser utilizado a técnica de remoção de *stopwords* que consiste em remover as palavras mais frequentes que na maioria das vezes não revelam informações relevantes para a construção do modelo ou, em alternativa e para o caso concreto da descoberta de dados pessoais, desenvolver um pequeno algoritmo para ensinar em que situações é que é importante reconhecer este tipo de palavras. Uma possível solução idealizada, mas não implementada consiste em construir um módulo de padrões com o modo padrão associando ao algoritmo edificado para reconhecimento de termos compostos por intermédio da leitura do modo dicionário e MaxEnt.

# Referências


[1] Assembleia Geral das Nações Unidas, Declaração Universal dos Direitos do Homem, Paris, France: ONU, 1948.

[2] Diário da República, 1ª Série A, n.º 247, Lei n.º 67/98, de 26 de outubro - Lei da Protecção de Dados Pessoais, Lisboa: Assembleia da República, 1998.

[3] Diário da República, Série I, n.º 86/1976, de 10 de abril, Constituição da República Portuguesa, Lisboa: Assembleia da República, 1976.

[4] Parlamento Europeu e do Conselho, Regulamento (UE) 2016/679, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados, Bruxelas: Jornal Oficial da União Europeia, 27 de abril de 2016.

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

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

[7] K. Peffers, T. Tuunanen, M. A. Rothenberger and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems,* vol. Vol. 24 No.3, 2007.

[8] N. Prat, I. Comyn-Wattiau and J. Akoka, "Artifact Evaluation in Information Systems Design-Science Research - a Holistic View," in *PACIS 2014 Proceedings. Paper 23*, 2014.

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

[10] V. Vaishnavi e B. Kuechler, “Design Science Research in Information Systems,” Georgia State University & University of Nevada, USA, 2013.

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

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

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

[14] V. Hristidis, L. Gravano e Y. Papakonstantinou, Efficient IR-Style Keyword Search over Relational Databases, Berlim, Germany: Proceedings of the 29th VLDB Conference, 2003.

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

[16] 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.

[17] J. Nothman, N. Ringland, W. Radford, T. Murphy e J. R. Curran, Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence 194, 2013.

[18] B. Hachey, W. Radford, J. Nothman, M. Honnibal e J. R. Curran, Evaluating Entity Linking with Wikipedia, Artificial Intelligence 194, 2013.

[19] N. Konstantinova, Review of Relation Extraction Methods: What Is New Out There?, Switzerland: Springer International Publishing, 2014.

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

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

[22] INESC-ID, “L2F INESC-ID,” [Online]. Disponível em: https://www.l2f.inesc-id.pt/. [Acedido em 26 07 2018].

[23] INESC-ID, “String - L2F's wiki,” 19 03 2015. [Online]. Disponível em: https://string.l2f.inesc-id.pt/w/index.php/Main_Page. [Acedido em 26 07 2018].

[24] 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.

[25] Armenak, “Sensitive Data Management: Data Discovery and Anonymization toolkit,” Github, 2017. [Online]. Disponível em: https://github.com/armenak/DataDefender. [Acedido em 06 11 2017].

[26] Apache Software Foundation, “Apache OpenNLP,” Apache, 2017. [Online]. Disponível em: https://opennlp.apache.org/. [Acedido em 29 11 2017].

[27] OpenNLP, “OpenNLP Documentation,” OpenNLP sourceforge, [Online]. Disponível em: http://opennlp.sourceforge.net/README.html. [Acedido em 02 03 2018].

[28] Diário da República, 1ª Série A, n.º 247, Lei n.º 68/98, de 26 de outubro - Determina a entidade que exerce as funções de instância nacional de controlo e a forma de nomeação dos representantes do Estado Português na instância comum de controlo, Lisboa: Assembleia da República, 1998.

[29] CNPD, 10 medidas para preparar a aplicação do regulamento Europeu para a proteção de dados, Lisboa: CNPD, 2016.

[30] 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.

[31] ISO, “International Organization for Standardization,” 2018. [Online]. Disponível em: https://www.iso.org/about-us.html. [Acedido em 23 07 2018].

[32] ISO, “ISO/IEC 27000:2018 - Information technology - Security techniques - Information security management systems - Overview and vocabulary,” ISO/IEC, Switzerland, 2018.

[33] ISO, “ISO/IEC 29100 - Information technology - Security techniques - Privacy framework,” ISO/IEC, Switzerland, 2011.

[34] ISO, “NP ISO/IEC 27001:2013 - Tecnologia de Informação - Técnicas de Segurança - Sistemas de Gestão de Segurança da Informação (Requisitos),” IPQ, Caparica, 2013.

[35] ISO, “ISO/IEC 270002:2005 - Information technology - Security techniques - Code of practice for information security management,” ISO/IEC, Switzerland, 2005.

[36] ISO, “ISO/IEC 29151:2017 - Information technology - Security techniques - Code of practice for personally identifiable information protection,” ISO/IEC, Switzerland, 2017.

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

[38] ISO, “ISO/IEC 29134:2017 - Information technology - Security techniques - Guidelines for privacy impact assessment,” ISO/IEC, Switzerland, 2017.

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

[40] Grupo de Trabalho do Artigo 29.º, “WP 248 rev.01 - Orientações relativas à Avaliação de Impacto sobre a Proteção de Dados,” 4 outubro 2017. [Online]. Disponível em: http://ec.europa.eu/justice/data-protection/index_en.htm. [Acedido em 29 04 2017].

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

[42] ISO, “NP ISO 31000:2013 - Gestão do Risco - Principios e Linhas de Orientação,” IPQ, Caparica, 2013.

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

[44] Project Mangement Institute, A Guide to the Project Management Body of Knowledge, 5 ed., Pennsylvania: PMI, 2013.

[45] LinkConsulting, “EAPY Data Governance - User Manual,” LinkConsulting, Lisboa, 2017.

[46] ISACA, “COBIT 5: Enabling Processes,” ISACA, USA, 2012.

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

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

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

[50] RedGlue, “REDATASENSE - Find sensitive data with Machine Learning techniques,” Redatasense, 2018. [Online]. Disponível em: http://redglue.eu/redatasensewp/. [Acedido em 26 09 2018].

[51] RedGlue, “Sensitive Data Discovery tool,” GitHub, Inc, 2018. [Online]. Disponível em: https://github.com/redglue/redsense. [Acedido em 20 04 2018].

[52] Tutorialkart, “tutorialkart.com,” 2017. [Online]. Disponível em: https://www.tutorialkart.com/. [Acedido em 20 02 2018].

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

[54] Apache Software Foundation, “The Apache Software Foundation,” 2018. [Online]. Disponível em: https://www.apache.org/dyn/closer.cgi/opennlp/opennlp-1.8.4/apache-opennlp-1.8.4-bin.tar.gz. [Acedido em 28 09 2018].

[55] Apache Software Foundation, “Apache OpenNLP Tools 1.9.0 API,” Apache Software Foundation, 2018. [Online]. Disponível em: https://opennlp.apache.org/docs/1.9.0/apidocs/opennlp-tools/index.html. [Acedido em 05 10 2018].

[56] Oracle, “Package java.util.regex,” Java™ Platform Standard Ed.8, 2018. [Online]. Disponível em: https://docs.oracle.com/javase/8/docs/api/java/util/regex/package-summary.html. [Acedido em 18 06 2018].

[57] Juristas, “Código Penal - Atualizado de acordo com a Lei n.º 44/2018, de 09 de Agosto,” 2018. [Online]. Disponível em: http://www.codigopenal.pt/. [Acedido em 28 03 2018].

[58] A. I. Carvalho e R. Lafuente, “Central de Dados,” 26 06 2018. [Online]. Disponível em: http://centraldedados.pt/codigos_postais/. [Acedido em 04 07 2018].

[59] Ministério das Finanças, “Modelo 3 - Declaração de Rendimentos - IRS,” Autoridade Tributárias e Aduaneira, Lisboa, 2018.

[60] APF, “Identidade e orientação sexual,” Associação para o Planeamento da Familia, [Online]. Disponível em: http://www.apf.pt/sexualidade/identidade-e-orientacao-sexual. [Acedido em 20 03 2018].

[61] DGAEP, “Tabela de habilitações literárias,” Direção-Geral da Admistração e do Emrpego Público, Lisboa, https://www.dgaep.gov.pt/upload/homepage/Noticias/LVCR/TAB_LVCR_HABILITACOES.pdf.

[62] IRN, “Nomes próprios de cidadãos portugueses nos últimos 3 anos,” Instituto dos Registos e do Notariado, 09 11 2017. [Online]. Disponível em: http://www.irn.mj.pt/sections/irn/a_registral/registos-centrais/docs-da-nacionalidade/vocabulos-admitidos-e/. [Acedido em 20 03 2018].

[63] Diário da República, 1ª Série A, n.º 136, Decreto Regulamentar n.º 76/2007, de 17 de julho, Lisboa: Conselho de Ministros, 2007.

[64] Orphanet, “O portal para as doenças raras e os medicamentos órfãos,” Orphanet versão 5.15.0, 2018. [Online]. Disponível em: https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=PT&search=Disease_Search_List. [Acedido em 20 03 2018].

[65] Apache Software Foundation, “Class RegexNameFinder,” 2017. [Online]. Disponível em: https://opennlp.apache.org/docs/1.7.2/apidocs/opennlp-tools/opennlp/tools/namefind/RegexNameFinder.html. [Acedido em 18 04 2018].

[66] Sourceforge.net, “Online regex tester,” [Online]. Disponível em: http://myregexp.com/. [Acedido em 02 09 2018].

[67] Jan Goyvaerts, “Regex Tutorial,” 2018. [Online]. Disponível em: https://www.regular-expressions.info/tutorial.html. [Acedido em 24 07 2018].

[68] AMA - Agência para a Modernização Administrativa, “Validação de Número de Documento do Cartão Cidadão,” 26 Janeiro 2009. [Online]. Disponível em: https://www.autenticacao.gov.pt/documents/10179/11463/Valida%C3%A7%C3%A3o+de+N%C3%BAmero+de+Documento+do+Cart%C3%A3o+de+Cidad%C3%A3o/0dbc446b-3718-41e5-b982-551a72f8b8a8. [Acedido em 17 07 2018].

[69] Dreispt, “Validators_pt,” [Online]. Disponível em: https://gist.github.com/dreispt/024dd11c160af58268e2b44019080bbf. [Acedido em 14 05 2018].

[70] Segurança Social, “Seguranças Social - Direta,” República Portuguesa, 06 02 2012. [Online]. Disponível em: http://www.seg-social.pt/pedido-de-niss1. [Acedido em 20 03 2018].

[71] Network Working Group, “RFC 5322 - Internet Message Format,” STD1 - Internet Official Protocol Standards, 10 2008. [Online]. Disponível em: https://tools.ietf.org/html/rfc5322#section-3.4. [Acedido em 02 03 2018].

[72] Diário da República, 1ª Série A, n.º 38, “Decreto-Lei n,º 45/2005, de 23 de fevereiro - Disposições relativas ao modelo comunitário de Carta de Condução,” Conselho de Ministros, Lisboa, 2005.

[73] ECBS, “IBAN: INTERNATIONAL BANK ACCOUNT NUMBER,” European Committee for Banking Standards, Bruxelas, 2003.

[74] ISO, “ISO 8601:2004 - Data elements and interchange formats - Information interchange - Representation of dates and times,” ISO/IEC, Switzerland, 2004.

[75] Portugal-a-programar, “Função checkdigit do Cartão do Cidadão,” Invision Community, [Online]. Disponível em: https://www.portugal-a-programar.pt/forums/topic/73408-fun%C3%A7%C3%A3o-checkdigit-do-cart%C3%A3o-do-cidad%C3%A3o/. [Acedido em 21 04 2018].

[76] Webdados, “Validação de Número de Contribuinte Português,” 19 12 2016. [Online]. Disponível em: https://www.webdados.pt/2014/08/validacao-de-nif-portugues-em-php/. [Acedido em 21 04 2018].

[77] W3Schools, “w3schools.com - The World's Largest Web Developer Site,” W3Schools, 2018. [Online]. Disponível em: https://www.w3schools.com/js/default.asp. [Acedido em 14 07 2018].

[78] Wikipedia, “Precision and recall,” Wikimedia Foundation, Inc., 03 10 2017. [Online]. Disponível em: https://en.wikipedia.org/wiki/Precision_and_recall. [Acedido em 04 12 2017].

[79] Responsáveis TI Marinha, Interviewee, *Estudo de Caso: Marinha - Validação dos resultados.* [Entrevista]. 01 10 2018.

[80] Link, “Edoclink - White paper,” 2017. [Online]. Disponível em: http://www.linkconsulting.com/edoclink/wp-content/uploads/sites/22/2017/10/Edoclink_WhitePaper.pdf. [Acedido em 08 10 2018].

[81] Responsaveis Edoclink Link Consulting, Interviewee, *Estudo de Caso: Link Consulting - Sessão de apresentação dos resultados.* [Entrevista]. 09 10 2018.

[82] P. Bernard, Foundations of ITIL 2011 Edition, Van Haren Publishing, 2012.

[83] P. Domingos, A revolução do Algoritmo Mestre, Barcarena: Manuscrito, 2017.

[84] NIST SP 800-39, “Managing Information Security Risk: Organization, Mission, and Information System View,” NIST - National Institute of Standards and Technology, USA, 2011.

[85] ISACA, "COBIT Process Assessment Model (PAM): Using COBIT 5," ISACA, USA, 2013.

[86] ISACA, "COBIT Self-assessment Guide: Using COBIT 5," ISACA, USA, 2013.

# Apêndices

# Apêndice A Relação entre o COBIT 5 e o RGPD

O quadro abaixo procura levantar um conjunto de possíveis questões a fazer para permitir refletir sobre as preocupações a ter em conta sobre a privacidade dos dados, nomeadamente quando se tem a responsabilidade de tratar dados pessoais. Todas as questões formuladas estão relacionadas com um ou mais artigos do RGPD e com áreas de domínio consideradas comuns entre o RGPD e o COBIT 5.

Tabela A.1: Possíveis questões relacionando o COBIT 5 com o RGPD

| **Áreas de Domínio ISACA – COBIT 5** | | **Nr** | **Possíveis questões para garantir o cumprimento do RGPD** | **Artigos RGPD** |
|---|---|---|---|---|
| Escolha e Consentimento | 1.1 | 1 | Existem políticas de privacidade definidas? | 6.º (1) |
| | .2 | 2 | Existe uma política definida para os procedimentos de recolha e tratamento dos dados? | 6.º (1) |
| | .3 | 3 | Os consentimentos obtidos estão documentados e guardados adequadamente? | 7.º (1) (2) |
| | .4 | 4 | Recolhe informações de crianças com menos de 16 anos de idade? Se sim, e processos implementados para recolher o consentimento dos pais conforme exigido pelo RGPD? | 8.º (1) |
| | .5 | 5 | Se "SIM", estão definidas, criadas e implementadas políticas para recolher e armazenar o consentimento dos pais. | 8.º (1) |
| Especificação do propósito (legitimidade) e Limitação da utilização | 2.1 | 6 | As políticas de privacidade e segurança preveem que se recolha apenas os dados pessoais que são adequados, pertinentes e limitados ao que é necessário relativamente às finalidades para as quais os dados são tratados ("minimização dos dados")? | 5.º (1) |
| | .2 | 7 | As políticas de privacidade e segurança garantem que o tratamento de dados pessoais seja lícito e necessário de acordo com as finalidades para as quais os dados foram recolhidos? | 6.º (1b) |
| | .3 | 8 | Existem políticas de privacidade e segurança para garantir que qualquer tratamento futuro que não o pretendido inicialmente seja revisto e tratado adequadamente antes de tal uso (obtenção de consentimento adicional)? | 6.º (4a) |
| | .4 | 9 | A organização efetua tratamento de dados pessoais relacionados com condenações penais e infrações? | 10.º |
| | .5 | 10 | Se "SIM". As políticas de privacidade e segurança garantem que o tratamento e/ou uso de dados pessoais de condenação criminal estejam sujeitos a um controlo exclusivo das autoridades oficiais ou autorizado pela União ou Estado-Membro? | 10.º |
| | .6 | 11 | Está contemplado e previsto as situações em que o direito de contestar não se aplica, e tem implementado procedimentos de apoio apropriados para aplicar nestas situações? | 22.º (2) |

| | | | | |
|---|---|---|---|---|
| | .7 | 12 | As responsabilidades de trabalho do responsável pela proteção de dados tem em consideração os riscos existentes para os dados pessoais e os riscos de possíveis violações associados aos titulares de dados, de modo a que a finalidade e a limitação do tratamento dos dados possam ser considerados os mais adequados? | 39.º (2) |
| Informações Pessoais e<br>Ciclo de vida de informações<br>Confidenciais | 3.1 | 13 | As políticas de privacidade e segurança preveem conservar os dados pessoais apenas durante o período necessário para as finalidades para os quais são tratados (finalidade legal, interesse público, científicos e de investigação histórica)? ("limitação da conservação") | 5.º (1) (e) |
| | .2 | 14 | Recolhe e trata dados pessoais das categorias especiais, condenações criminais? | 6.º (4) |
| | .3 | 15 | Se "SIM". Possui mecanismos para ajudar a identificar se os referidos dados possam ser utilizados para fins diferentes do propósito original para os quais foram recolhidos? | 6.º (4) |
| | .4 | 16 | As políticas de privacidade e segurança definidas e aplicadas permitem assegurar e determinar se os seguintes tipos de dados pessoais são recolhidos (e/ou tratados) sob exceções relevantes fornecidas no RGPD ou se tal tratamento precisa ser proibido?<br>a) dados reveladores de origem racial ou étnica, opiniões políticas, crenças religiosas ou filosóficas ou filiação sindical.<br>b) dados genéticos.<br>c) dados biométricos com o propósito de identificar unicamente uma pessoa natural.<br>d) dados referentes à saúde.<br>e) dados relativos à vida sexual ou orientação sexual de uma pessoa singular | 9.º (1) |
| | .5 | 17 | As políticas de privacidade e segurança permitem que os titulares de dados possam ser removidos de utilizar os seus dados pessoais para efeitos de comercialização direta sempre que solicitarem a oposição ao tratamento de tais comunicações? | 21.º (3) |
| | .6 | 18 | Existe forma de conseguir relacionar a legitimidade dos dados pessoais do titular dos dados? | 21.º (1) |
| | .7 | 19 | Existem e estão implementados mecanismos técnicos e organizacionais para assegurar que, por defeito, apenas sejam tratados os dados pessoais que forem necessários para a finalidade especifica do tratamento? | 25.º (2) |
| | .8 | 20 | As políticas de privacidade e segurança prevê a realização de procedimentos, quando necessário, para determinar se o tratamento é realizado de acordo com a avaliação de impacto da proteção de dados, sempre que o risco representado pelas operações de tratamento for alterado? | 35.º (11) |
| | .9 | 21 | O tratamento de dados pessoais destina-se para fins de investigação científica ou histórica ou para fins estatísticos? Se "SIM", está a prever derrogações: | 89.º (2) |
| | .10 | 22 | Aos direitos de acesso do titular dos dados? (Artigo 15) | 89.º (2) |
| | .11 | 23 | Ao direito de retificação? (Artigo 16) | 89.º (2) |
| | .12 | 24 | Ao direito à limitação do tratamento? (Artigo 18) | 89.º (2) |
| | .13 | 25 | Ao direito de oposição? (Artigo 21) | 89.º (2) |

| | | | | |
|---|---|---|---|---|
| Precisão e Qualidade | 4.1 | 26 | Possui mecanismos para conseguir garantir que os dados pessoais tratados são precisos e atualizados, conforme necessário? para corrigir erros de dados pessoais sem atraso? | 5.º (1) (d)<br>16.º |
| | .2 | 27 | O mecanismo adotado permite que os dados sejam apagados ou retificados sem demora? (pretende medir o nível de exatidão dos dados) | 5.º (1) (d)<br>16.º |
| Transparência e Notificação | 5.1 | 28 | Garante que os dados pessoais são recolhidos de acordo uma finalidade claramente específica e legítima? isto é, não são utilizados para fins de tratamento diferentes dos indicados e que são tratados de forma justa, transparente e em conformidade com os requisitos legais aplicáveis? | 5.º (1) (b) |
| | .2 | 29 | Possui mecanismos que permitem comunicar/informar aos titulares de dados os seus direitos e responder de uma forma concisa, transparente e utilizando uma linguagem clara, onde e como os seus dados pessoais são tratados? | 12.º (1) |
| | .3 | 30 | Está habilitado a fornecer, no momento em que os dados pessoais são recolhidos, todos os elementos de informação necessários, tais como os direitos dos titular dos dados; como restringir a sua utilização; como proceder para retirar o consentimento, ...; bem como para garantir um tratamento justo e transparente? | 13.º (1) (2)<br>14.º (2)<br>21.º (4) |
| | .4 | 31 | Existem políticas de privacidade e segurança para informar os titulares dos dados sobre as garantias aplicadas quando os dados pessoais são transferidos para um país terceiro ou uma organização internacional? | 15.º (2) |
| Participação Individual | 6.1 | 32 | As politicas de privacidade e segurança permitem que os titulares de dados retirem o consentimento de utilização dos seus dados pessoais a qualquer momento (incluindo dados pessoais usados em parceria com outros controladores), desde que a cancelamento não resulte em violações legais sobre as quais informou os titulares dos dados? | 7.º (3)<br>26.º (3) |
| | .2 | 33 | As politicas de privacidade e segurança preveem a parceria com outros controladores conjuntos para garantir que um titular de dados cuja identidade foi verificada possa exercer seus direitos de?<br>- Acesso a;<br>- Informação sobre;<br>- Correções para;<br>- Eliminação / destruição (apagamento) de; - Restrições a dados pessoais associados em conformidade com o tempo, custos e formato dos requisitos de entrega de informações exigidos pelo RGPD?<br>Incluindo processos para fornecer razões documentadas para negar solicitações? | 12.º (2) (3) (4) (5) (6)<br>14.º (3)<br>16.º<br>17.º<br>21.º (1)<br>26.º (3) |
| | .3 | 34 | Existe capacidade para permitir dar reposta ao titular dos dados a confirmação de que os dados pessoais que lhe digam respeito são ou não objeto de tratamento? (aplica-se para os casos que utilizem parcerias com outros responsáveis pelo tratamento de dados) | 15.º (1)<br>18.º |
| | .4 | 35 | Quando solicitado pelo titular, em média quanto tempo demora a responder as seguintes informações?<br>A Finalidade do tratamento de dados;<br>As Categorias dos dados pessoais em questão;<br>Os Destinatários (partilha de contratos);<br>Períodos de conservação dos dados (retenção);<br>Direitos de exclusão e registro de reclamações;<br>Capacidade de restringir o tratamento de dados pessoais quando viável e legal, com avisos quando | 15.º (1)<br>18.º |

| | | | | |
|---|---|---|---|---|
| | | | as restrições são levantadas;<br>(incluiu igualmente o artigo 26.º (3)) | |
| | .5 | 36 | As políticas de privacidade e segurança contemplam o fornecimento de cópias dos dados pessoais que não se destinam ao exercício de funções de interesse público ou por autoridades oficiais, cuja cópia deve ser tratada a pedido da pessoa em causa, sem prejuízo; entregue num formato digital comumente usado, junto com cópias adicionais conforme solicitado; por uma taxa razoável, onde a taxa é baseada em custos administrativos reais?<br>« Inclui o art.º 15 (3) » | 20.º (3)<br>15.º (3) |
| | .6 | 37 | As politicas de segurança e privacidade permitem que os titulares de dados contestem o uso de seus dados pessoais para efeitos de comercialização direta, incluindo a definição de perfis, incluindo aqueles que resultam em decisões ou circunstâncias que afetam significativamente o titular dos dados? | 22.º (1)<br>21.º (2) |
| | .7 | 38 | As politicas de segurança e privacidade permite que os titulares de dados contactem o encarregado da proteção de dados para qualquer questão relacionada ao tratamento dos seus dados pessoais ou ao exercício dos direitos que lhe são conferidos pelo RGPD? | 38.º (4) |
| Responsabilização<br>(Accountability) | 7.1 | 39 | | 5.º |
| | .2 | 40 | As politicas de privacidade e segurança detalham:<br>a) a base legal aceitável para o tratamento de dados pessoais, conforme exigido pela lei; e<br>b) o procedimento para o tratamento de dados para outros fins é compatível com a finalidade da recolha original dos dados pessoais (tendo em conta o contexto em que os dados pessoais foram recolhidos e em particular a relação entre os titulares de dados e empresa) | 6.º (1) (3) (4) |
| | .3 | 41 | A recolha dos dados pessoais são recolhidos presencialmente? | 14.º (1) |
| | .4 | 42 | Se "Não". As políticas de privacidade e segurança garantem que é fornecido ao titular dos dados as seguintes informações:<br>a) Identidade e os contactos do responsável pelo tratamento;<br>b) Contato de qualquer oficial de proteção de dados;<br>c) Finalidades e fundamentos jurídicos para o tratamento;<br>d) Documentação das categorias dos dados pessoais em causa;<br>e) Destinatários (ou categorias de destinatários) dos dados pessoais, se houver;<br>f) Registos da intenção de transferir dados pessoais para um destinatário num país terceiro ou organização internacional, quando aplicável;<br>g) Existência ou ausência de uma decisão de adequação pela Comissão; e<br>h) Referências às garantias e os meios de obter uma cópia das mesmas | 14.º (1) |
| | .5 | 43 | O responsável pelo tratamento ou subcontratante encontra-se estabelecido fora da União? Se "SIM", devem designar por escrito um representante que pertença à União | 3.º (2)<br>27.º (1) |
| | .6 | 44 | Possui um código de conduta aprovado e publicado? (pode ser utilizado como elemento para demonstrar o cumprimento das obrigações do RGPD) | 24.º (3)<br>32.º (3)<br>40.º (2) |

| | | | | |
|---|---|---|---|---|
| | .7 | 45 | As politicas de privacidade e segurança prevê a repartição de responsabilidades, a sensibilização e formação do pessoal implicado nas operações de tratamento de dados, e as auditorias correspondentes? (de modo a cumprir com os requisitos de RGPD) | 39.º (1) (b) |
| | .8 | 46 | As politicas de privacidade e segurança prevê que o responsável consulte a autoridade de controlo apropriada e forneça informações (incluindo responsabilidades do controlador, propósitos e meios de tratamento, garantias implementadas, proteções contra violações à privacidade, etc. .), conforme exigido pelo RGPD, antes do tratamento, sempre que a AIPD indique que o tratamento resultaria num elevado risco na ausência das medidas tomadas para mitigar o risco? | 36.º (1) (3) |
| | .9 | 47 | As politicas de privacidade e segurança detalham:<br>a) Requisitos para o estabelecimento das responsabilidades do encarregado da proteção de dados;<br>b) As tarefas para as quais o encarregado pela proteção de dados será responsável, em conformidade com o RGPD; e<br>c) As medidas em vigor para assegurar que a(s) pessoa(s) que desempenham a função tenham a devida qualificação e conhecimento dos requisitos legais para a proteção de dados? | 37.º (1) (2) (3) (4) (5) (6) |
| | .10 | 48 | O responsável pelo tratamento é responsável pelos princípios relativos ao tratamento de dados pessoais pelo que deverá garantir que:<br>a) A autoridade do EPD seja compreendida e reconhecida;<br>b) O Encarregado da proteção de dados esteja envolvido em todas as questões relativas a dados pessoais;<br>c) Todos os executivos, incluindo os de mais alto nível, não promovem apenas suporte para o EPD, mas também fornecem os recursos necessários (incluindo conhecimento e formação) para cumprir as responsabilidades da posição;<br>d) O EPD não seja penalizado por desempenhar funções ou manter a necessária confidencialidade; e<br>e) o EPD também é encarregado de outras responsabilidades além daquelas de ser oficial de proteção de dados, conforme apropriado e razoável dado o ambiente de negócios? | 38.º (1) (2) (3) (5) (6)<br><br>5.º(2) |
| Garantias de Segurança | 8.1 | 49 | Existem políticas de privacidade e segurança para assegurar:<br>a) Quais são os mecanismos a serem implementados para garantir a proteção dos dados pessoais, incluindo proteções contra tratamento não autorizado ou ilegal e contra perda, destruição ou dano acidental; e<br>b) Que medidas técnicas e/ou organizacionais sejam aplicadas quando dados pessoais são usados para finalidades diferentes daquelas para as quais os dados foram inicialmente recolhidos (por exemplo, cifragem, controlos de acesso, pseudonimização, políticas, formação, registros, etc. )? | 5.º (1)<br>6.º (4)<br>24.º (2) |
| | .2 | 50 | Existem políticas de privacidade e segurança para implementar as medidas técnicas e organizacionais apropriadas que garantam um nível de segurança apropriado ao risco de violações pessoais, incluindo (conforme apropriado):<br>a) Utilização de pseudonimização e/ou cifragem;<br>b) Procedimentos para estabelecer a confidencialidade, integridade, disponibilidade e resiliência dos sistemas e serviços de tratamento, backup e recuperação de dados; e<br>c) Testes regulares dos controlos de segurança associados? | 32.º (1) |
| | .3 | 51 | Existem políticas de privacidade e segurança para avaliar a probabilidade de violações de privacidade aos titulares de dados no caso de: | 32.º (2) |

| | | | | |
|---|---|---|---|---|
| | | | a) Acesso não autorizado, partilha ou divulgação dos dados pessoais; e<br>b) Destruição, perda ou alteração não autorizada ou acidental? | |
| | .4 | 52 | Existem políticas de privacidade e segurança para obter autorização, quando aplicável, da autoridade de controlo competente para garantias apropriadas sobre os dados pessoais, por meio de:<br>a) Cláusulas contratuais entre os responsáveis pelo tratamento ou subcontratantes e os responsáveis pelo tratamento, subcontratantes ou destinatários dos dados pessoais no país terceiro ou organização internacional; ou<br>b) Disposições a inserir nos acordos administrativos entre as autoridades ou organismos públicos que incluam direitos vinculativos e eficazes para os titulares de dados? | 46.º (3) |
| Monitorização,<br>Métricas e Relatórios | 9.1 | 53 | Existem políticas de privacidade e segurança para fornecer relatórios aos titulares de dados (em horários especificados, mediante solicitação, conforme apropriado, e refletindo todos os componentes exigidos pelo RGPD), sempre que se verifique:<br>a) Violação de dados pessoais;<br>b) Qualquer retificação ou apagamento dos dados pessoais ou limitação do tratamento;<br>c) Portabilidade dos dados. | 17.º<br>19.º<br>20.º (1) (2)<br>34.º (2) |
| | .2 | 54 | As politicas de privacidade e segurança devem ainda comunicar aos titulares dos dados, sempre que solicitado pelo o mesmo:<br>a) Relatórios mostrando o conteúdo dos dados pessoais associados à finalidade;<br>b) Relatórios mostrando os dados pessoais que foram partilhados com terceiros, incluindo as razões para tal partilha;<br>c) Cópias digitais completas dos dados pessoais transmitidos diretamente para outro controlador de dados, em apoio aos requisitos de portabilidade de dados | 17.º<br>19.º<br>20.º (1) (2)<br>34.º (2) |
| | .3 | 55 | Existem políticas de privacidade e segurança para conservar os registos de todas as atividades relacionadas com o tratamento de dados, constando a seguinte informação:<br>a) Nome e contatos do responsável e a lista dos vários elementos que estão envolvidos no processo;<br>b) Finalidade do tratamento;<br>c) Descrição das categorias;<br>d) As categorias de destinatários a quem os dados pessoais foram divulgados;<br>e) Se possível, prazos previstos para o apagamento das diferentes categorias;<br>f) Se possível, Descrição geral das medidas técnicas e organizativas no domínio da segurança. | 30.º |
| | .4 | 56 | Existem políticas de privacidade e segurança que acionam as AIPD de acordo com o RGPD, geralmente nas seguintes situações:<br>a) ao processar dados pessoais usando novas tecnologias e sistemas;<br>b) quando solicitado ou determinado como necessário pelo EPD;<br>c) quando necessário para o tratamento automatizado e tomada de decisão, incluindo sistemas de perfis;<br>d) ao processar grandes quantidades de categorias especiais de dados pessoais ou dados pessoais relacionados a condenações e infrações penais;<br>e) sempre que os sistemas monitorarem áreas de acesso público em larga escala;<br>f) sempre que as operações de tratamento envolvam dados pessoais; e<br>g) sempre que monitorar os dados sujeitos? | 35.º (1) (2) (3) (4) |

| | | | | |
|---|---|---|---|---|
| | .5 | 57 | Existem políticas de privacidade e segurança que regem o conteúdo requerido nos relatórios da AIPD, conforme determinado pelo RGPD, incluindo:<br>a) Uma descrição sistemática das operações de tratamento, finalidades e interesse legítimo do responsável;<br>b) Uma avaliação da necessidade e proporcionalidade das operações de tratamento em relação aos propósitos;<br>c) Uma avaliação dos riscos para os direitos e liberdades relacionados aos dados; e<br>d) Uma avaliação das medidas necessárias para mitigar os riscos (incluindo, por exemplo, garantias, medidas de segurança e mecanismos para proteger os dados pessoais e demonstrar conformidade com o RGPD)? | 35.º (7) |
| | .6 | 58 | Existem políticas de privacidade e segurança para fornecer às autoridades de controlo apropriadas informações de contato do responsável pela proteção de dados e relatórios sobre violações da segurança de dados pessoais, refletindo todos os componentes exigidos pelo RGPD? | 37.º (7)<br>33.º (5) |
| | .7 | 59 | Existem políticas de privacidade e segurança que especificam as funções e/ou tarefas cuja responsabilidade é do EPD, incluindo:<br>a) Monitorizar a conformidade (incluindo a conformidade dos subcontratantes de dados com os requisitos);<br>b) Formação do pessoal;<br>c) Assegurar o desempenho de auditorias de conformidade com a privacidade;<br>d) Assegurar o desempenho das AIPD;<br>e) Cooperar com as autoridades de controlo; e<br>f) Fornecer quaisquer outras informações, conforme apropriado. | 39.º (1) |
| | .8 | 60 | Existem políticas de privacidade e segurança para manter um registro de atividades de tratamento envolvendo dados pessoais que incluem:<br>a) O nome e detalhes do contato da organização e, quando aplicável, tais detalhes para qualquer representante e/ou EPD;<br>b) A finalidade para o tratamento de dados pessoais;<br>c) Uma descrição das categorias de titulares de dados e das categorias de dados pessoais envolvidos no tratamento;<br>d) As categorias de destinatários a quem os dados pessoais foram ou serão divulgados, incluindo destinatários em países terceiros ou organizações internacionais;<br>e) Transferências de dados pessoais para um país terceiro ou uma organização internacional, quando aplicável, incluindo a identificação desse país terceiro ou organização internacional e a documentação das garantias associadas; e<br>f) Quando possível, os prazos estabelecidos para o apagamento das diferentes categorias de dados? | 47.º (2) |
| Prevenção de violações | 10.1 | 61 | Existem políticas de privacidade e segurança que especifiquem como determinar se o tratamento de dados pessoais (incluindo o tratamento para finalidades diferentes daquelas para as quais os dados pessoais foram inicialmente recolhidos) é legal porque:<br>a) O sujeito de dados associado forneceu consentimento explícito;<br>b) É necessário cumprir o contrato com a pessoa em causa;<br>c) É necessário para cumprimento legal;<br>d) É necessário proteger os interesses vitais da pessoa em causa ou de outras pessoas singulares;<br>e) É necessário para o interesse público ou para o exercício da autoridade oficial; ou | 6.º (1) (4) |

| | | | | |
|---|---|---|---|---|
| | | | f) É necessário apoiar interesses legítimos da sua empresa que não infrinjam os direitos do titular dos dados? | |
| | .2 | 62 | Existem políticas de privacidade e segurança que especificam possíveis consequências para os titulares de dados para qualquer tratamento posterior? | 6.º (1) (4) |
| | .3 | 63 | Existem políticas de privacidade e segurança para garantir que os titulares de dados exerçam seus direitos para alterar a forma de como os seus dados pessoais são utilizados? solicitar cópias de dados pessoais; e/ou exercer outros direitos ao abrigo do RGPD desde que não afetem negativamente os direitos e liberdades de outros. | 15.º (4)<br>20.º (4) |
| | .4 | 64 | Existem políticas de privacidade e segurança que descrevam como garantir que as decisões relativas aos titulares de dados não devam ser feitas com base em categorias especiais de dados pessoais, a menos que as garantias específicas tenham sido implementadas? | 22.º (4) |
| | .5 | 65 | Existem políticas de privacidade e segurança que especificam os direitos dos titulares de dados de solicitar a remoção dos seus dados pessoais de um sistema automatizado e do perfil em situações que podem resultar em efeitos legais adversos ou violações a eles, e também para especificar as ações (e métodos de contato associados) que sua empresa pode tomar para obter as visões dos titulares de dados (ou seus representantes legais) sobre o tratamento pretendido? | 35.º (9) |
| | .6 | 66 | Existem políticas de privacidade e segurança para fornecer medidas técnicas e organizacionais que garantam o respeito ao princípio da minimização de dados, conforme detalhado no RGPD, sempre que dados pessoais forem usados para: a) o interesse público; b) fins de pesquisa científica ou histórica; ou c) fins estatísticos? | 89.º (1) |
| | .7 | 67 | Existem políticas de privacidade e segurança relacionadas ao uso de informações de igrejas e associações religiosas e comunidades? | 91.º (1) |
| Gestão de Terceiros/Fornecedores | 11.1 | 68 | Tem definido políticas de privacidade e segurança na gestão dos contratos com terceiros/fornecedores de modo a garantir que não utiliza subcontratantes, a menos que:<br>a) forneçam garantias suficientes e comprovem que implementaram técnicas e organizacionais apropriadas (físicas e administrativas); medidas e controlos que cumpram os requisitos do RGPD e apoiem os direitos dos titulares de dados;<br>b) o responsável pela proteção de dados ou equivalente fornecer autorização por escrito para usar o subcontratante; e<br>c) o subcontratante concorde contratualmente em notificar sua organização sempre que adicionar ou remover outros subcontratantes? | 28.º (1) (2) |
| | .2 | 69 | Tem definido políticas de privacidade e segurança na gestão dos contratos com terceiros/fornecedores para especificar o tipo de contrato (em cópia impressa e/ou digital) ou outro ato legal sob a lei ou de Estado-Membro, vinculando o subcontratante em relação à sua empresa, que estabelece:<br>a) o objeto e a duração do tratamento;<br>b) a natureza e finalidade do tratamento;<br>c) os tipos de dados pessoais, categorias de titulares dos dados e as oito categorias de obrigações exigidas no âmbito do RGPD;<br>d) a obrigação de notificar sua organização sobre quaisquer violações, mudanças antecipadas nas obrigações e os direitos de sua empresa de verificar tais requisitos? | 28.º (3) (9) |

| | | | | |
|---|---|---|---|---|
| | .3 | 70 | Tem definido políticas de privacidade e segurança na gestão dos contratos com terceiros/fornecedores que detalham as ações que os subcontratantes devem executar caso envolvam outros subcontratantes a realizar atividades de tratamento específicas que façam parte das atividades que a organização contratou o subcontratante? executar e garantir que tal subcontratação inclua os mesmos requisitos com os quais o subcontratante concordou dentro do contrato que eles têm com a sua organização, incluindo a verificação da prova de implementação e a existência desses requisitos, além da prova de conformidade com qualquer código de conduta envolvido e os requisitos de certificação, e deixa claro que o subcontratante que trabalha diretamente com sua organização permanece totalmente responsável pelo desempenho de todas as obrigações? | 28.º (4) (5) |
| | .4 | 71 | Tem definido políticas de privacidade e segurança na gestão dos contratos com terceiros/fornecedores que especificam as etapas que sua organização deve executar para garantir que pessoas físicas ajam sob a autoridade de sua organização (e aquelas dos seus subcontratantes que tenham acesso a dados pessoais)? siga:<br>a) todas as políticas e procedimentos de dados pessoais;<br>b) instruções fornecidas por sua organização ou pelo subcontratante para o qual as pessoas físicas trabalham; e<br>c) todas as regras associadas e quaisquer requisitos estabelecidos pela União ou Lei do Estado-Membro | 29.º<br>32.º (4) |
| Gestão de Violações | 12.1 | 72 | Tem políticas documentadas de violação de dados pessoais (e procedimentos de apoio) que incluem requisitos para:<br>a) notificar as autoridades de controlo apropriadas sobre a violação em tempo útil e com as razões fornecidas para quaisquer atrasos;<br>b) notificar os titulares de violação de alto risco (conforme definido pelo RGPD) no prazo máximo de 72 horas após a descoberta de uma violação, se for determinado (seguindo procedimentos documentados para realizar a análise de risco de violações) que a violação de dados pessoais resultará em violações à privacidade aos titulares de dados associados; e c) incluir todos os itens necessários dentro do aviso, conforme exigido pelo RGPD? | 33.º (1) (2) (3) (4)<br>34.º (1) (3) |
| Privacidade e<br>Segurança por Omissão | 13.1 | 73 | Tem políticas documentadas e aplicadas (e procedimentos de apoio) para criar proteções de segurança e privacidade em todo o ciclo de vida de processos automatizados de tomada de decisão envolvendo dados pessoais; salvaguardar os direitos, liberdades e interesses legítimos da pessoa em causa; para permitir a intervenção humana por sua empresa (como o controlador); permitir que os titulares de dados associados incluam os seus pontos de vista sobre as decisões associadas; e permitir que os titulares dos dados contestem as decisões? | 22.º (3) |
| | .2 | 74 | Tem políticas documentadas e aplicadas (e procedimentos de apoio) para avaliar (ou de outra forma ter em conta) os riscos associados à natureza, âmbito, contexto e propósitos do tratamento de dados pessoais e a probabilidade e gravidade associadas a violações para dados? e projetar, construir e implementar controles de segurança e privacidade técnicos, administrativos e físicos apropriados, apoiados por princípios de privacidade documentados (por exemplo, os Princípios de Privacidade da ISACA e/ou padrões de privacidade da ISO, etc.), para mitigar adequadamente esses violações. na medida do possível, em conformidade com o RGPD e para proteger os direitos dos titulares de dados? | 24.º (1) |

| | | | | |
|---|---|---|---|---|
| | .3 | 75 | Tem politicas documentadas e aplicadas (e procedimentos de apoio) para aplicar medidas técnicas e organizativas para assegurar que, por omissão, só sejam tratados os dados pessoais que forem necessários para cada finalidade específica do tratamento? | 25.º (1) (2) |
| Responsabilidades do controlador,<br>Automatização,<br>Tomada de decisão e<br>Proteção de dados por omissão | 14.1 | 76 | Tem políticas documentadas e aplicadas de dados pessoais (e procedimentos de apoio) para apoiar o tratamento legal que:<br>a) ocorre sem o consentimento do sujeito de dados e / ou aviso, e em situações específicas em que o titular dos dados tenha objetado; mas<br>b) apoia propósitos legítimos e documentados de interesse público, autoridade oficial, pesquisa científica e pesquisa histórica? | 6.º (1)<br>21.º (6) |
| | .2 | 77 | Tem documentado e aplicado políticas de transferência de dados pessoais (e procedimentos de apoio) que incluem as etapas a seguir para transferir dados pessoais para um país terceiro ou para uma organização internacional somente após certas condições terem sido validadas de acordo com todos os requisitos de transferência progressiva? | 44.º<br>45.º<br>46.º (1)<br>48.º<br>47.º (1) (2) |
| | .3 | 78 | Tem políticas documentadas e aplicadas (e procedimentos de apoio) para contatar a autoridade de controlo apropriada, usando o mecanismo de consistência estabelecido e associado, para aprovar regras corporativas vinculantes para garantir que elas sejam legalmente vinculantes; incluir todas as proteções de dados necessárias e apropriadas; são consistentemente aplicados; fornecer todos os direitos legalmente exigidos dos titulares de dados; e cumprir os requisitos RGPD? | 47.º (1) (2) |
| | .4 | 79 | Tem políticas documentadas e aplicadas de transferência de dados pessoais (e procedimentos de suporte), que na ausência de decisões de adequação e regras corporativas vinculantes, fornecem um processo para transferir dados pessoais somente sob uma das sete condições listadas no RGPD? e somente se:<br>a) a transferência não for repetitiva;<br>b) diz respeito apenas a um número limitado de titulares de dados;<br>c) é necessário para os propósitos ou para persuadir interesses legítimos prosseguidos pelo responsável pelo tratamento, cujos objetivos não são sobrepostos pelos interesses, direitos e liberdades do titular dos dados;<br>d) implementou garantias para mitigar adequadamente os riscos de segurança associados a todas as circunstâncias que envolvem a transferência de dados; e<br>e) forneceu informações à autoridade de controlo apropriada sobre a transferência? | 49.º (1) |
| | .5 | 80 | Tem políticas de segurança de dados documentadas e aplicadas (e procedimentos de apoio) para implementar garantias de transferência de dados pessoais, conforme aplicável para cada situação, sob acordos legais com autoridades públicas; regras corporativas obrigatórias; cláusulas-tipo da autoridade de controlo aplicável; códigos de conduta aprovados; ou mecanismos de certificação aprovados, conforme detalhado no RGPD? | 46.º (2) |

# Apêndice B Questões de avaliação de ameaças e/ou vulnerabilidades

Todas as questões elaboradas foram pensadas de forma a obter uma resposta binária (sim ou não) e resumem os pontos considerados mais críticos na segurança da privacidade de dados na vertente da segurança tecnológica.

As questões efetuadas foram maioritariamente baseadas no RGPD e na Resolução de Conselho de Ministros n.º 41/2018 em detrimento das normalizações porque (1) os diplomas atendem às normalizações (que são documentos de referência para as boas práticas) e (2) devido a apresentarem um vínculo de obrigatoriedade no seu cumprimento sujeitando a coimas para quem não o cumprir.

**Controlo do acesso a sistemas e aplicações (Autenticação)**

1. O processo de autenticação é iniciado mantido de uma forma segura (como por exemplo, o uso de palavra-passe)?
2. As credenciais de inicio de sessão garantem a integridade e confidencialidade na comunicação entre as entidades tecnológicas? Se utiliza *hash* ou cifra.
3. A palavra-passe dos utilizadores é complexa e a sua dimensão é no mínimo de 9 caracteres?
4. A palavra-passe dos administradores é complexa e a sua dimensão é no mínimo de 13 caracteres?
5. Tem implementado um sistema de autenticação de multifactor? Isto é, se a palavra-passe é utilizada com a combinação de um outro fator (*token*, *sms*,…)
6. Possui bloqueio automático das estações de trabalho (computadores) após 5 minutos, no máximo?

**Politicas de privilégios mínimos**

7. Os perfis de utilizadores estão criados de acordo com os privilégios mínimos (principio da necessidade de executar)?
8. As contas de acesso dos utilizadores estão de acordo com as suas funções e os privilégios de acesso estão associados às funções?
9. O acesso ao tratamento de dados pessoais é efetuado através de estações de trabalho (computadores) específicos para o efeito?
10. Estes computadores possuem configurações estáticas de configuração de rede e estão ligados a portas especificas de acesso?

**Monitorização**

11. Existem registos (*logs*) de todas as atividades referentes ao tratamento de dados (acesso, alteração, transferência, remoção)? Nomeadamente informação de quem acedeu, de onde acedeu, quando acedeu, a que dados acedeu e que ação foi efetuada? (ponderar apenas para os dados do artigo 9.º).

12. Todas as contas dos utilizadores são auditáveis regularmente com uma periodicidade mínima de 180 dias para os utilizadores normais e 90 para os administradores?
13. Possui mecanismos automáticos de alarmística para as contas de utilizadores sem atividade?
14. Todos os registos (*logs*) de atividade são armazenados apenas em modo de leitura e são assinados digitalmente (de modo a preservar a sua integridade)?
15. São armazenados os registos de todos os acessos? Incluindo as tentativas falhadas?
16. Os registos de atividade (*logs*) armazenam o endereço de acesso (IP e porto), *hostname*, *hash* da conta do utilizador, registo do dia e hora ao segundo (*timestamp*) e a ação efetuada?

**Comunicações de dados**

17. Os dados são transmitidos de uma forma segura? Mascarados, anonimizados ou cifrados?
18. Os dados pessoais armazenados (backups) encontram-se cifrados e assinados digitalmente?

**Conhecimento de segurança e dos regulamentos**

19. Conhece ou possui um inventário rigoroso com todos os sistemas relacionados com a recolha e tratamento de dados pessoais?
20. Tem o vínculo contratual ou termo de consentimento por parte do titular de dados de todos os dados pessoais que recolhe e trata?
21. Os elementos responsáveis pelo processo de recolha e tratamento dos dados pessoais possuem formação adequada às suas funções (nomeadamente sobre segurança de proteção e privacidade de dados)?
22. Os responsáveis pelo processo de recolha e tratamento de dados pessoais estão sensibilizados e credenciados (qualificados) com as politicas e regulamentos em vigor?
23. Possui um código de conduta elaborado, aprovado e divulgado?
24. O acesso aos ativos da infraestrutura tecnológica está devidamente isolado e protegido?
25. Existe monitorização e controlo dos espaços físicos dos ativos da infraestrutura?

**Utilizadores**

26. O acesso aos repositórios de dados pessoais está limitado apenas aos elementos responsáveis pelo seu tratamento?
27. O acesso aos repositórios é efetuado por intermédio de aplicação (*front-end*)?
28. Existem elementos com capacidade de aceder diretamente aos repositórios sem ser por intermédio de aplicação dedicada para o efeito?
29. Se respondeu afirmativamente à questão anterior, existem mecanismos de registos (*logs*) de atividade?
30. Tem o controlo positivo do perfil dos utilizadores que acedem à base de dados? administradores, utilizadores?

# Apêndice C Medidas de proteção

As medidas de proteção apresentadas são baseadas nas melhores práticas de segurança da área tecnologia e na privacidade de dados, apresentadas pelas normalizações e recomendações internacionais. Alerta-se para o facto que as medidas sugeridas são de caracter abrangente (geral) devendo ser feito um exercício ponderado à realidade de cada organização e de acordo com a dimensão do volume e criticidade do tipo de dados pessoais que trata.

**Controlo do acesso a sistemas e aplicações**

1. Independentemente do sistema operativo utilizado e/ou aplicação, deve garantir de todos os acessos são baseados em protocolos de autenticação. O mais simples e económico é baseado num segredo partilhado (através do uso de palavra-passe).
2. Mediante o nível protocolar deverá escolher o mecanismo mais adequado para garantir a integridade e confidencialidade das comunicações. Possíveis soluções: PGP, PEM; SSH, HTTP, IMAPS, POPS, TLS/SSL; IPSEC; GSM, WEP, Bluetooth).
3. Deverá garantir e forçar nas politicas de gestão para que os utilizadores sejam obrigados a ter uma senha que seja complexa com um mínimo de 9 caracteres. Para ser fácil de memorizar recomenda-se a utilização de uma pequena frase com alterações de alguns caracteres para símbolos e números.
4. Deverá garantir e forçar nas politicas de gestão para que os administradores possuem uma senha complexa com um mínimo de 13 caracteres. Para ser fácil de memorizar recomenda-se a utilização de uma pequena frase com alterações de alguns caracteres para símbolos e números.
5. Recomenda-se a implementação de um sistema multimétodo, no mínimo, para os administradores de rede e sistemas. Note-se que os elementos de prova usados nos protocolos de autenticação são baseados (1) no que se sabe – ex: senha, (2) o que se possui – ex: cartão ou telemóvel e (3) o que se é – ex: impressão digital. Combine mais do que um elemento.
6. Para além de incentivar o bloqueio por parte dos utilizadores quando abandonem a estação de trabalho, force as politicas de segurança da infraestrutura para bloquear automaticamente após 5 minutos, no máximo.

**Politicas de privilégios mínimos**

7. Crie e utilize contas de utilizador institucionais em detrimento de contas pessoais e defina os privilégios de acordo com a função e a necessidade de executar.
8. Edifique uma matriz de funções e o nível de privilegio e implemente na politicas da sua organização. Procure garantir se um determinado utilizador desempenhe duas funções distintas que possua duas contas e utiliza as contas de acordo com a função. Principio dos privilégios mínimos.
9. Procure garantir que os utilizadores com acesso ao tratamento de dados pessoais possuam uma estação de trabalho própria e que não seja partilhada com outro utilizador.

10. Garanta que todos os computadores que tenham acesso e tratem dados pessoais possuem configurações estáticas de configuração de rede (IP fixo), configuração da porta de rede dedicada exclusivamente à máquina (através do *MAC Address*).

**Monitorização**

11. Procure ter um sistema centralizado de registos *(SIEM – Security Information and Event Management)* e no caso de tratar dados pessoais de acordo com o artigo 9.º configure o sistema e/ou aplicações para promoverem todos os registos dos eventos relacionados com o tratamento de dados (acesso, alteração, transferência, remoção). Nomeadamente informação de quem acedeu, de onde acedeu, quando acedeu, a que dados acedeu e que ação foi efetuada.
12. Garanta que possui mecanismos para conseguir auditar regularmente todas as contas dos utilizadores que tratem dados pessoais são com uma periodicidade mínima de 180 dias para os utilizadores normais e 90 para os administradores?
13. Avalie se todos os sistemas e aplicações possuem mecanismos automáticos de alarmística para as contas de utilizadores sem atividade. Tenha especialmente atenção aos sistemas legados mais antigos. Em caso afirmativo, avalie bem os riscos e pondere evoluir para um sistema atual.
14. Force o registo dos eventos de atividades relacionadas com o tratamento de dados apenas ao modo de leitura e garanta que apenas sejam acessíveis por intermédio de dupla autenticação (duas pessoas distintas) para os assinar digitalmente. A periodicidade de deverá ser reajustada de acordo com o nível de criticidade dos dados.
15. Defina as politicas de gestão dos controlos de acesso para se registar todas os processos de autenticação, incluindo as tentativas falhadas.
16. Deve procurar garantir que os registos das atividades (*logs*) relacionadas com o tratamento de dados pessoais contenham, preferencialmente, o endereço de acesso (IP e porto), *hostname*, *hash* da conta do utilizador, registo do dia e hora ao segundo e a ação efetuada. Quando se trate de dados do artigo 9.º esta ação é mandatória.

**Comunicações de dados**

17. Para qualquer transferência de dados pessoais deve garantir a utilização de mecanismos de segurança adequados, sendo que (1) para dados críticos (referentes ao artigo 9.º do RGPD) os dados devem estar cifrados; (2) para os restantes dados poderá adotar outros mecanismos como o mascaramento, anonimização e/ou pseudonimização. Em complemento, recomenda-se a utilização de *VPN*[21] para garantir a identidade correta do remetente e destinatário.
18. Tendo em conta a complexidade e custo desta medida, o sistema de *backup* deve, no mínimo, estar segregado logicamente, minimizando o seu acesso a um número mínimo de utilizadores. Recomendando-se que (1) apenas os sistemas que tratam dados pessoais críticos estejam devidamente cifrados; (2) em relação aos restantes dados pessoais, deve haver registos de todas

[21] *Virtual Private Network*

as atividades. Independentemente da solução deverá, no mínimo, avaliar os registos de *logs* mensalmente. Faça uma avaliação de impacto para conhecer o risco.

**Conhecimento de segurança e dos regulamentos**

19. Promova um inventário com todos os sistemas de informação que utiliza diferenciando os que são destinados exclusivamente para o tratamento de dados pessoais, classifique a criticidade dos dados pessoais, enumere o número de utilizadores com acesso aos dados e analise o fluxo do processo de tratamento dos dados pessoais.
20. Garanta que possui o vínculo contratual ou termo de consentimento por parte do titular de dados de todos os dados pessoais que trata e relacione-o com o sistema e/ou aplicações responsáveis pelo tratamento de dados.
21. Dê formação e crie incentivos para todos os colaboradores com responsabilidade no processo de recolha e tratamento dos dados pessoais (nomeadamente sobre segurança de proteção e privacidade de dados). Promova regularmente, treino e avaliação de situações que pretende testar (procedimentos sobre o processo de tratamento de dados institucionalizado).
22. Para além das qualificações necessária para os elementos com funções no processo de tratamento de dados, procure promover ações internas sobre as politicas e regulamentos institucionalizados na organização.
23. Elabore rapidamente o código de conduta e envolva a direção no processo para aprovar e divulgar quer pelos colaboradores quer para os titulares dos dados. Procure ser o mais transparente na divulgação do código de conduta de modo a garantir uma maior lealdade e compromisso.
24. Procure ter espaços técnicos dedicados para acomodar os ativos da infraestrutura tecnológica, limitando o acesso a um número restrito de elementos, devidamente identificados. Evite ter equipamentos ativos de rede acessíveis (fisicamente e logicamente) a qualquer elemento.
25. Implemente um sistema de controlo de acessos aos espaços físicos (técnicos e de trabalho) e preferencialmente, instale um sistema de vídeo como complemento à monitorização dos espaços técnicos.

**Utilizadores**

26. Condicione o acesso físico e logico aos repositórios de dados (com especial enfase aos dados pessoais) e garanta mecanismos de controlo de acesso e monitorização dos espaços.
27. Procure na medida do possível implementar aplicações (*front-end*) para aceder aos repositórios garantindo que a aplicações promova autenticações seguras e registe todas os eventos das atividades realizadas. Adicionalmente, deverá assegurar mecanismos de proteção de acordo com o tipo de utilizador e função, ou seja, é nesta componente que são aplicadas as medidas de proteção de acordo com a área de negócio (pseudonização, anonimização e/ou cifragem).
28. Anule de imediato acessos diretos. No caso de não ser possível e/ou viável restrinja ao máximo o número de utilizadores que possam aceder diretamente aos repositórios sem ser por intermédio de aplicação dedicada para o efeito. Tenha permanentemente presente esta situação e implemente o mais rápido possível uma forma de monitorizar e ter todos os registos das ações/atividades desenvolvidas (seja auditável).

29. Está perante uma situação de alto risco, principalmente se trata dados pessoais considerados sensíveis. Implemente o mais rápido possível uma forma de monitorizar e ter todos os registos das ações/atividades desenvolvidas para tornar o sistema auditável.
30. No caso de já possuir sistemas e/ou aplicações que façam o controlo automatizo parametrize e crie regras para promover alertas regulares. No caso do controlo ser manual não facilite, mantenha permanentemente atualizado os perfis de utilizadores e sempre que se verificar uma alteração de funções atualize de imediato.

# Apêndice D Guião de entrevista e questionário

Guião de questões para entrevista semiestruturada:

1. Relativamente ao propósito (descoberta de dados pessoais), mediante os resultados obtidos e apresentados na demonstração como é que classifica o protótipo?

2. Quais é que são os aspetos que considera/classifica como positivos (mais vantajosos)?

3. Na sua opinião, que modificações ou alterações podem ser feitas para melhorar o protótipo?

4. Considera que a utilização do protótipo possa ser uma mais valia na organização
   4.1. Como acelerador para detetar dados pessoais?
   4.2. Para detetar desvios de comportamentos nos processos implementados?
   4.3. Como complemento ao nível de conformidade do RGPD?
   4.4. Para ajudar na avaliação do grau de privacidade sobre os dados pessoais?

5. Com base nos conhecimentos sobre a segurança sobre a privacidade de dados e na gestão de informação como é que classifica o produto demonstrado?

Mini-questionário:

1. A sua opinião do protótipo, quanto à sua utilização, isto é facilidade de executar?
   Simples ou complexo. Escala de 1 a 5.
2. Satisfação dos tempos de resposta no processo de descoberta de dados?
   Lento ou rápido. 1 a 5.
3. Agregação e apresentação dos resultados?
   Difícil ou fácil. 1 a 5.
4. Facilidade de compreender/entender o conceito?
   Difícil ou fácil. 1 a 5.
5. Satisfação dos resultados em relação ao expectável?
   Nada ou muito. 1 a 5.
6. Utilizaria ou recomendaria o protótipo na sua organização?
   Sim ou não
7. Qual é a sua apreciação global?
   Muito má a excelente. 1 a 5

# Apêndice E Resumo de entrevistas da Marinha

## CMG EMT Manuel da Costa Honorato Data: 2018-08-30

**(Diretor da Direção de Análise e Gestão da Informação - DAGI)**

Parece-me obvio que o trabalho tem muito interesse e que tudo o que for feito nesta área é fundamental. A primeira dificuldade que a organização tem sentido neste domínio é claramente na questão da pesquisa, identificação e na inventariação dos dados pessoais, pelo que é fundamental trabalhos como este e outros relacionados com o tema. Mas temos, também, de ter a humildade de perceber que estamos perante um trabalho que tem vindo a ser desenvolvido num ambiente académico pelo que não é espectável ser um trabalho perfeito e na maioria das vezes é um trabalho realizado num ambiente totalmente novo e desconhecido e quando estamos a trabalhar no desconhecido temos sempre hipótese de reavaliar, reaprender em todo este processo mas considero que o trabalho tem todas as premissas necessárias para iniciar um processo (até porque um processo que não se esgota) e com base nos resultados devem ser feitas várias aproximações sucessivas. A curva de aprendizagem não respeita uma equação matemática bem definida, não se está perante um modelo bem defindo e linear, logo é necessário de se fazer várias iterações e otimizações. Para isto e com a experiência é claro que vamos começar a ficar mais críticos e com melhor capacidade de analisar outras variáveis. Portanto, não tenho dúvidas quanto ao propósito definido que estamos a trabalhar numa área muito útil para a Marinha, para o Estado como um todo, diria mesmo para o país. É uma área sempre de inovação constante e permanente, notei que existe muita investigação por detrás do protótipo desenvolvido, parece-me que existe igualmente inovação, pelo que preenche todos os requisitos de um trabalho de índole académico, mas obviamente há também uma aplicação prática e isso acresce valor. Não estamos perante um trabalho puramente académico de investigação, mas sim perante um trabalho de engenharia e reconheço valor e mérito por isso.

Pela demonstração e pelas prontas respostas verifico um aspeto muito positivo que foi a abordagem de aproximações sucessivas seguida, ou seja, foi procurado fazer uma abordagem de aprendizagem e não de conceitos pré-concebidos, e isso é bom. Pelo que o aspeto mais vantajoso, obviamente para além do produto é a possibilidade de aprendizagem e do crescimento, incluindo do teu próprio conhecimento… escolheste seguir esse caminho e isso é bom.

Quanto às alterações, tenho receio da classificação que foi atribuída em relação aos falsos positivos porque poderemos ter uma dimensão elevada que pode ofuscar, criar demasiado ruído e poderá limitar a eficácia da própria descoberta de dados pessoais, mas lá está, estamos no universo académico a desenvolver uma prova de conceito (protótipo) e apenas se pode melhorar com várias iterações (aproximações sucessivas) dos próprios modelos para refinar a sua eficácia e cobertura. Parece-me que temos de estar atentos e eventualmente revisitar este processo dos falsos positivos porque poderá descredibilizar o produto, mas creio que ainda mais importante que os falsos positivos são os falsos negativos do que não detetar verdadeiros positivos, mas parece-me que o risco seja menos porque a

aproximação que se está a fazer é através de termos isolados (palavras). Mais que uma alteração, diria que é para estar atento e ir corrigindo situações pontuais

O melhoramento que aponto tem haver com a questão de se poder promover uma capacidade incremental de autoaprendizagem acho que será muito interessante, acredito não seja um processo fácil devido à sua complexidade e até eventualmente requeria uma outra dissertação de mestrado, daí a importância de definir bem o âmbito e militar a investigação. Outro aspeto, mas esse poderá ser identificado com trabalho futuro seria estender para além da componente meramente tecnológica.

Dentro do projeto, há aspetos a melhorar, sim. Mas também há aspetos de possível melhoria com base na experiência e conhecimento adquirido e há aspetos que claramente estão fora do âmbito do trabalho.

Quanto ao interesse para a Marinha, isso é inquestionável… claramente que sim. Até diria que não se restringe apenas ao nível da Marinha, mas sim ao nível do Estado. Neste momento é difícil quantificar a exatidão de todos os sistemas e aplicações que estão em produção no estado em geral que tratam dados pessoais, mas seguramente que são demasiados, pelo que estes tipos de iniciativas são necessários, é pena não existir capacidade de recursos para continuar a promover o trabalho.

Reconheço que o protótipo cumpre o seu papel como acelerador de identificação dos dados pessoais, já para a identificação de potenciais desvios parece que sim, mas irá obrigar a ter um modelo de referência. Quanto a conseguir ter uma conformidade com o RGPD já tenho mais dúvidas porque não se restringe apenas à componente tecnológica, temos sempre uma componente mais subjetiva, ou seja, irá ajudar, mas apenas faz meio caminho, mas lá está, mais vale ter meio caminho feito e acelerar o processo do que não ter nada e andar passo a passo.

Não concordo muito com o termo impacto, creio que não é muito feliz porque o termo impacto normalmente está patente quando alguma coisa acontece e projeta para o futuro, mas se estivermos perante o grau de privacidade, aí sim, o protótipo certamente irá ajudar.

Se o objetivo foi ter uma dissertação que não fosse apenas restringida a investigação, mas que tivesse uma vertente aplicacional, diria que sim, o objetivo foi conseguido. Assim como não cabe num produto académico ter resultados próximos de requisitos industriais e os resultados finais parecem-me já muito bons.

Finalmente, considero que o produto apresentado tem potencial, embora seja académico já apresenta um bom nível. É pena não termos sinergias dentro das Forças Armadas ou Estado que possam pegar nestas pequenas iniciativas e que de uma forma colaborativo possam dar continuidade a produtos que são desenvolvidos no universo académico e transformá-lo num produto industrial. Que tem potencial, sim, tem. E se for para dar continuidade, deverá ser envolvido vários intervenientes *(stakeholders)*, ter uma equipa multidisciplinar porque ninguém faz nada sozinho e teremos de alargar os requisitos aos ambientes não tecnológicos, uma melhoria na otimização e eficácia, capacidade de incluir processos de aproximações sucessivas no próprio sistema, robustez, usabilidade, resiliência e eventualmente capacidade de integrar com a componente do *business intelligence*.

Em suma, globalmente considero que o produto está muito positivo e apresenta um bom nível, verifico que os princípios estão presentes, que existe uma carga de investigação, de inovação e houve uma preocupação de índole prático (aplicacional).

## CMG EMA Luís Eduardo Moita Rodrigues Data: 2018-10-01

**(Diretor da Direção de Tecnologias de Informação e Comunicações - DITIC)**

Em relação ao protótipo propriamente dito diria que representa um pequeno passo, na boa direção, visto que permite de uma forma simples e com base num conjunto de critérios pré-definidos e normas que regulam esta matéria, fazer a descoberta de dados pessoais numa determinada infraestrutura tecnológica/rede de dados, dando a conhecer os tipos de dados que uma organização poderá ter espalhados. Permite também fazer pesquisas do particular para o geral, ou seja, permite realizar pesquisas em apenas um computador, vários computadores de um domínio, podendo alargar a vários domínios com pesquisas em rede a um servidor ou vários servidores (repositórios), ou seja, é uma ferramenta que permite de acordo com as necessidades escalar o universo da pesquisa. Classifico este protótipo de bom, face aos resultados apresentados e à sua potencial aplicabilidade/utilização dentro da organização. Como ferramenta de descoberta de dados pessoais.

Considero que o protótipo poderá ser uma boa ajuda numa fase inicial em que organização não tem a noção ou consciência onde de facto existem dados passíveis de serem considerados dados pessoais. Na realidade, acredito, que, ao contrário do que que deveria acontecer (i.e., dados pessoais corretamente armazenados e acesso aos mesmos controlado com medidas de proteção adequadas), muitos dados pessoais estão espalhados sem qualquer tipo de controlo e acessíveis a muito utilizadores sem necessidade de acesso a essa informação. Creio que tal não acontece por maldade, mas sim fruto de exigências do dia a dia e, muitas vezes, como medida facilitadora na realização de determinadas tarefas. Claro que existem colaborados que estão mais conscientes e despertos para este tipo de situações, mas outros apenas guardam informação porque sim, sem razão aparente. Neste sentido diria que o protótipo é sem duvida uma boa ajuda.

Sendo um protótipo, relevo como principal vantagem/ atributo a sua simplicidade, pelo facto de possibilitar, de forma simples e rápida, efetuar uma avaliação sobre uma determinada amostra e obter um retrato/ conhecimento situacional sobre eventuais dados pessoais existentes. No que respeita aos modelos, é importante ter consciência do nível de exatidão. É um facto que quanto mais rica for a base de conhecimento, melhor serão os resultados, mas diria que com o tempo vai ser possível maturar os modelos e consequentemente melhorar a sua eficácia e desempenho na descoberta e classificação de dados pessoais, exatamente a função para o qual o mesmo foi desenvolvido.

Um aspeto a ter cuidado, e mais importante que os próprios modelos, é a tradução dos modelos para algoritmos/ fórmulas de cálculo, e a sua posterior validação com recurso a uma base de dados conhecida. Um pequeno erro neste processo pode deitar tudo a perder e retirar toda a credibilidade ao protótipo. É, pois, essencial compreender bem os algoritmos, o que produzem, e validar o mesmo com recurso a uma amostra conhecida, como já referido, verificando se os resultados são os esperados de

acordo com o modelo implementado. Só assim se conseguirá dar credibilidade ao protótipo e modelos adotados.

Outro aspeto a ter em consideração, é o facto da variabilidade linguística em texto corrido ser demasiada complexa e que nem todas as pessoas escrevem da mesma maneira. Contudo, é sempre possível melhorar o desempenho do protótipo neste particular. Por exemplo, pode-se melhorar a cobertura para a morada através da utilização de um marcador no código postal, que por norma apresenta bons resultados de cobertura e precisão, e que permite identificar com rigor o local, caso a informação encontrada seja medíocre.

Como melhoramento, embora não estejamos perante um produto comercial e sabendo que para um protótipo isso não seja uma preocupação, importa conseguir um bom nível de usabilidade da eventual ferramenta que venha a ser desenvolvida na sequência dos resultados alcançados com este protótipo. Neste particular, é importante ter uma interface gráfica apelativa e funcional que permita com apenas alguns cliques, selecionar os vários critérios (modo de consulta, tipos de dados a descobrir e local a procurar) e de uma forma rápida e intuitiva colocar a aplicação a correr. O mesmo se aplica para a interpretação e apresentação dos resultados, devendo permitir navegar de uma forma interativa entre os vários níveis de visualização dos resultados.

Considero que o protótipo pode ser uma mais valia para a organização e que contribui, em maior ou menor grau, para todos os pontos referidos na questão quatro. Contudo, o nível de ambição e objetivos a alcançar devem ser os superiormente definidos pela organização, no qual o EPD desempenhará um papel fundamental. Mas sim, uma das primeiras preocupações que se deve ter é conhecer aonde é que existem dados pessoais no domínio de marinha e, em particular, nas redes locais. Esta ferramenta pode, de facto, e numa fase inicial, contribuir para a obtenção desse conhecimento, saber como e onde estamos neste particular, para a partir dai definir uma estratégia concertada que permita corrigir as inconformidades encontradas, indo ao encontro do RCM n.º 41/2018.

Acrescentaria ainda, que não basta detetar comportamentos desviantes é preciso atacá-los reeducando as pessoas, algo que pode ser feito com a adoção de medidas tão simples, como por exemplo a criação de doutrina/normas internas para regular a gestão da informação, em particular a informação relativa a dados pessoais, normalizar processos e procedimentos, envolver e responsabilizar as pessoas …, ou seja, medidas que, no seu conjunto, concorrem para uma mudança de comportamentos a nível individual e para a mudança da cultura organizacional neste domínio.

Finalizo como comecei: a utilização deste protótipo permitirá à organização fazer um RX e aferir o estado atual no que respeita à localização de dados pessoais e, por essa via, tomar consciência da realidade. Independentemente dos resultados que forem obtidos, estes traduzirão, com a precisão assegurada pelo modelo adotado, o estado atual, sendo este o ponto de partida para se definir uma estratégia que vise assegurar a proteção, salvaguarda e privacidade dos dados pessoais. É certo que ainda existe um longo caminho a percorrer, inclusive na melhoria deste protótipo, contudo, este é um primeiro passo para “agarrar” este processo. O protótipo, no imediato e por si só, já teve o mérito de chamar a atenção para esta questão, demonstrar algumas fragilidades da organização neste domínio,

que implicarão, estou certo, a adoção de medidas corretivas a vários níveis de modo a assegurar a conformidade com a legislação em vigor relativa à proteção de dados pessoais.

## CMG SEP Jorge António Oliveira da Silva Rocha Data: 2018-09-20

**(STI - Chefe de Secretaria de Serviços Partilhados e Oficial de Segurança à Unidade)**

Relativamente à apresentação e ao trabalho desenvolvido até então considero excecionalmente útil, isto porque é necessário e fundamental para se começar a trabalhar no sentido de conseguir aplicar o RGPD e se não tivermos uma ferramenta dedicado ou similar ao protótipo desenvolvido, eventualmente mais evoluída, mas em boa verdade considero que o que já está feito é muito bom para se começar a perceber que tipo de dados existem nos computadores e iniciar o processo de analisar o que está menos bem.

Pelo que foi percebido, a ferramenta já identifica muitos dos dados que são passiveis de sofrerem correções. Nomeadamente, se tivermos acesso aos registos *(logs)* vai permitir fazer uma análise mais pormenorizada e a definir um plano para se começar a corrigir gradualmente.

Embora se esteja perante um protótipo de um trabalho académico, claramente que é um trabalho inicial (exploratório ou introdutório) e de acordo com os objetivos pretendidos e as funcionalidades elencadas diria que já apresenta um bom nível. Gostei particularmente dos graus de classificação que foi atribuída porque dá um alerta importante e considero que estão muito fiéis à interpretação do RGPD permitindo canalizar esforços para o que poderá ser mais importante.

Muito sinceramente, desconhecia os pormenores de como é que o trabalho estava a ser desenvolvido, mas por mim, embora seja um protótipo e não apresente características de um produto industrial, considero estarmos perante um produto em condições de exploração, permitindo a sua utilização pelos vários administradores do domínio de utilizador (ADU), o principio de funcionamento para começar a aplicar e explorar o tipo de dados que poderão estar localizados indevidamente nos vários computadores dos vários domínios de rede da Marinha.

Isto porque se todas as pessoas com responsabilidades no domínio de tratamento de dados, começarem a fazer este tipo de exercício nos seus domínios de utilização considero que a tarefa da missão de conformidade com regulamento seria bem mais facilitada.

Acima de tudo, considero que em termos práticos e, por experiência, só depois de se começar a explorar a ferramenta por terceiros é que realmente se poderá ter os melhores contributos para a melhorar. É sempre melhor produzir algum trabalho mesmo com menos recursos do que ficar à espera de algo (uma ferramenta) que cubra todas as necessidades , o que, infelizmente, não está ao nosso alcance de adquirir ou demorará demasiado tempo.

Eu sou suspeito porque as minhas raízes são de informático, mas pelo que vi, a ferramenta é perfeitamente usável, adorei ver que para a executar é com recurso a linha de comandos pelo que não apontaria este ponto como um aspeto a melhorar, mas seria interessante existir a capacidade de se

conseguir detetar documentos duplicados, não apenas pelo nome, mas sim com capacidade de comparar o seu conteúdo, isto porque, por mais que se sensibilize o pessoal que é para trabalharem diretamente com os sistemas colaborativos, elas tendem a trabalhar e a armazenar sempre uma cópia no seu próprio computador. Por exemplo, acontece inúmeras vezes uma pessoa iniciar a escrever um documento, mas por algum motivo é outra pessoa que vai dar continuidade a redigir esse mesmo documento, no caso da versão do documento não estar devidamente armazenado no repositório central ou eventualmente a versão que se encontra no repositório não é a mais recente perde-se imensa produtividade, parece que as pessoais não querem entender que trabalhar com sistemas partilhados serve para facilitar e não complicar, muito menos as controlar. Por algum motivo, se o documento está aberto por outro elemento, a pessoa recebe a informação de quem é que está a editar e pode bem perguntar se ainda precisa ou se o pode libertar.

Acredito que parte da falta de uma possível falha de conformidade do RGPD prende-se com esta questão de os colaboradores continuarem a armazenar continuamente a informação aonde não devem, deixando documentos que carecem de algum cuidado no seu manuseamento, completamente vulneráveis. Certamente, que muitos dos dados que iremos descobrir, com esta ferramenta ou outra qualquer, estarão relacionadas com situações deste género porque ao fazer este tipo de cópias estamos a contribuir para proliferar os dados e a realidade é que os computadores locais por mais mecanismos de segurança que possuem acabam por estar mais vulneráveis do que os repositórios dedicados para o efeito.

Como aspeto mais vantajoso, com já referi, identifico o nível de classificação atribuída pelos graus de importância, pois acho muito interessante este aspeto porque permite priorizar quais são os documentos que se deva investigar e analisar em primeiro lugar.

Sem dúvidas, estes tipos de iniciativas são muito benéficos para a organização, para além de se adquirir conhecimento específico na área permite que com uma simples ferramenta, se consiga iniciar o processo de análise sobre o tipo de dados existentes nos vários domínios do utilizador. Repare que neste momento não temos esta capacidade e claramente que, esta ferramenta poderá ser uma grande ajuda. Claro que estamos perante um protótipo e para a colocar em produção será necessário de ser melhorada ao nível de usabilidade, mas como referi, considero adequado a linha de comando. Por isso, quanto a mim, a ferramenta desenvolvida, pela demonstração prática e pelos resultados apresentados mesmo que, um modelo ou outro apresente uma eficácia e uma cobertura menos boa, diria que é suficiente para resolver parte da equação do RGPD, nomeadamente dar a conhecer o volume de dados pessoais existentes nos diversos computadores da Marinha.

Resumindo, para as necessidades mais básicas, o protótipo é mais do que suficiente porque normalmente as ferramentas industriais mais evoluídas fazem um conjunto de procedimentos que na prática acabam por não ser explorados porque na maioria das vezes não se adequam às nossas preocupações e necessidades. Este pequeno e simples protótipo parece-me, quanto a mim, uma maravilha porque vai alertar as pessoas para aspetos de utilização dos dados que nos dias de hoje, infelizmente. ainda não estão despertas.

**CFR SEP Júlio José Galo Penim Garcia** **Data: 2018-09-24**

**(DITIC - Chefe da Divisão de Sistemas de Software)**

Na minha perspetiva está bem patente que estamos perante uma ferramenta que traz uma mais valia a qualquer organização. Temos de ter a noção que a organização Marinha é demasiado grande (extensa) e com demasiada informação e na realidade a gestão de informação é algo ainda incipiente. Neste sentido é óbvio que devemos possuir ferramentas que, por um lado, permitam ter a capacidade de efetuar uma avaliação e, por outro, ter a capacidade de fazer algumas recomendações resultantes dessa mesma avaliação, pelo que creio que, claramente, a ferramenta que tem vindo a ser desenvolvida vem um pouco de encontro e responde a muitas destas necessidades.

Consciente que cada vez mais se produz demasiada informação, cada vez mais as organizações apresentam dificuldades na gestão de informação, e dos dados no geral, e na maioria dos casos dados sensíveis, sejam pessoais ou de outro nível de sensibilidade, é notório a necessidade eminente de ferramentas que facilitem a gestão dos dados, da informação e, eventualmente, do conhecimento.

Tenho acompanhado este processo desde o início e este tema foi muito bem adequado, quer em termos temporais, porque a questão do RGPD revelou-se um assunto prioritário nas organizações e está muito em voga, como também está muito adequado às nossas necessidades e ajustado à nossa realidade. Pois a Marinha, e repetindo-me, gera demasiado informação e nem sempre essa informação é devidamente administrada ou governada daí que existe a necessidade de gerir todo o universo de dados e conhecimento.

Creio que os resultados esperados de um produto deste género são muito similares à consequência dos processos de auditoria interna e de inspeção, ou seja, deve servir essencialmente para se identificar quais são as nossas vulnerabilidades e conhecer os aspetos que podemos melhorar e o resultado deste trabalho é claramente isto. Não estamos perante uma ferramenta perfeita (mas também convém não esquecer que estamos perante um trabalho académico) mas nada da vida é perfeito, contudo reconheço, e identifico, que os princípios chave estão presentes e é um muito bom ponto de partida para se começar a caracterizar o estado atual dos dados pessoais espalhados pela organização e o que poderá ser feito e melhorado.

O que considero que poderia ser alvo de melhoria prende-se com os vários módulos usados nas pesquisas, mas isso acabou por ficar um pouco condicionado devido à limitação do tempo da própria dissertação nomeadamente com a questão dos testes de avaliação do produto. Quanto a mim seria necessário ter mais tempo neste processo porque teria sido conveniente conseguir avaliar e confrontar diferentes tipos de sistemas, não só no aspeto da informação que se conhece e controla, para permitir aferir a precisão e cobertura, mas também, para se conseguir confrontar com diversos universos desconhecidos, teria sido muito interessante que tivesse sido possível expandir os testes a outros ambientes, para se ter amostras mais variadas. Digo isto porque o estudo de caso ficou limitado ao trabalho de uma secretaria do sector tecnológico e se fosse alargado a outros ambientes variados, certamente que haveria espaço para melhorar não só o motor de busca e pesquisa, como poder refinar os módulos dos vários modelos. Mas lá está, nem sempre tudo é perfeito e o tempo não é infinito, pelo

que considero que dentro da limitação do tempo, entre a investigação, o desenvolvimento e os testes realizados considero o produto final está bastante evoluído e com resultados muitos interessantes.

A forma como o processo foi conduzido e adaptado à dissertação considero que contribuiu de alguma forma para se ter um produto com utilidade futura, porque para além da descoberta de dados foi incorporado a componente de gestão dos dados e de certa forma a integração da RCM 41/2018. Com todos aqueles critérios e o levantamento que foi efetuado, o resultado para a Marinha é extremamente positivo (e todo esse esforço não é inteiramente visível). É fundamental caracterizar claramente cada sistema, e conseguir fazer a relação entre a realidade com o que é expectável, assim como se conseguir sumarizar um enorme volume de requisitos em algo mais objetivo e pragmático, não é propriamente fácil e tudo isso, claramente, é uma mais valia para o trabalho.

Relativamente a ser uma mais valia para organização, sem dúvidas que sim, para a Marinha e para qualquer outra organização. Questiono quais é que são, atualmente, as ferramentas ou sistemas que fazem uma avaliação a este nível e daí permitir saber quais são os dados pessoais que existem… de fonte aberta, não conheço nenhuma. É importante referir que internamente estamos a trabalhar num ambiente sensível à qual se adiciona a própria sensibilidade dos dados pessoais e, como referi, não houve tempo para explorar devidamente os dados relacionados com a saúde, os resultados toxicológicos, etc. Confesso que tenho algum receio sobre o tratamento desse tipo de dados e que estejam num nível de maturidade muito inferior aos que foram tratados nos estudos de caso e isto deve servir de base como considerações futuras.

Reforço, e volto a repetir o tema da gestão de informação, que considero ser a questão fundamental disto tudo. Em que estado é que estamos? Já em relação ao RGPD, fala-se muito, mas na realidade, e em termos concretos, isso resulta em muito pouco, temos muito pouca coisa onde possamos agarrar realmente para conseguir fazer este mapeamento. É verdade que estamos perante um trabalho académico e o resultado de uma dissertação funciona muito de um modo empírico, mas na realidade o resultado da tese traduz a devolução de dados concretos, reais e válidos que de alguma forma, e com base num argumento que está claramente identificado e é válido, em termos científicos considero que os resultados estão bem fundamentados.

Pelo que vi e do que conheço da ferramenta, pela forma como é intuitiva, de como é configurável, a própria facilidade de transportar e a modularidade no incremento de novos modelos, considero que é de todo conveniente incorporar nos trabalhos de auditoria e inspeções internas porque a componente do RGPD não se limita ao nível tecnológico e ao pessoal, mas é geral à organização. O produto desenvolvido com alguns, poucos ajustes, claramente que desempenha um papel preponderante para dar a conhecer aos responsáveis onde é que se encontram, ou estão espalhados, os dados pessoais.

**CTEN FZ Filipe da Rocha Rei** **Data: 2018-08-30**

**(DAGI - Chefe da Secção de Doutrina e de Gestão da Informação)**

Em relação à ferramenta, protótipo ou não protótipo, considero que são sempre fundamentais, ainda mais na era digital e na sua segurança. Portanto, a ferramenta considero que está boa, mas vou segregar os comentários em duas partes: (1) a primeira relaciona com a descoberta, que na minha opinião é claramente o elemento principal *(core business)* e que se deva apostar mais; e uma segunda (2) a componente de governação, que considero como um adicional *(add on)* e que tem de ser muito bem refletido para minimizar os enviesamentos que carecem de outras abordagens.

Focando na componente de descoberta, considero que para um protótipo já está muito boa e como aspeto a melhorar considero que deve ser muito bem pensado a taxonomia de classificação utilizada, pois deve estar muito bem definida para não haver dúvidas e a formula da classificação de risco, como é matemática, também deve ser bem refletida porque é sobre ela que vai permitir ser desenhado o plano de mitigação e devemos evitar que se siga o caminho errado.

É certo que estamos perante uma ferramenta para promover o alerta, mas rapidamente pode ser utilizada como uma ferramenta de comando e controlo, e como ferramenta de comando e controlo é essencial que os critérios de análise de risco não sejam alterados no tempo, só assim é que se consegue medir a evolução do nível de maturidade da organização. Em suma, como prova de conceito, está ótima, mas para ser implementado e entrar em produção, carece que seja efetuado uma reflexão cuidada e organizacional da classificação do grau de risco envolvendo várias partes interessadas no processo (ter uma abrangência de diversas áreas).

Outro melhoramento muito interessante seria possuir capacidade de autoaprendizagem, ser capaz de alimentar os modelos de uma forma semiautomática, ou seja, há medida que vai descobrindo os vários tipos de dados, o responsável ao validar determinados dados esses seriam incrementados permitindo enriquecer os modelos à medida da sua utilização (habilitar que seja evolutiva no tempo e melhorar gradualmente o nível de precisão e cobertura).

Refinar o processo das ambiguidades e a capacidade de reconhecer termos compostos, nomeadamente a questão das proposições e artigos definidos e as várias formas de se poder identificar determinados termos, mas este processo claramente é um processo interativo e evolutivo no tempo pelo que como primeira abordagem está muito bom e vejo que tem muito potencial, claramente que este deve ser o caminho a seguir.

Sem dúvida alguma que será uma mais valia para a organização, como referi, como primeira abordagem e como prova de conceito, o protótipo está bom e creio que de certa maneira, não tão direta, já responde às primeiras necessidades da organização e poderá ser utilizada como alerta de utilização indevida dos dados. Até porque, considero eu a situação mais critica que organização tem de enfrentar são com os dados não estruturados, pois de certa forma os dados estruturados estão devidamente controlados, são mais restritos no seu acesso e com politicas bem definidas, já os dados não estruturados é muito mais difícil controlar positivamente o que os vários colaboradores possuem nos computadores organizações, isto já para não falar dos computadores pessoais (portáteis) que são utilizados para fins profissionais.

Um melhoramento que pode ser pensado e feito tem haver com a duplicação, não só nos repositórios como nos computadores de trabalho. Questiono se a organização tem conhecimento qual é o volume de dados duplicados existentes nos seus repositórios e quanto é esta informação está a custar em termos de espaço útil do disco? Não acredito que se saiba…O mais importante é avaliar o conteúdo dos documentos porque é muito comum isto ocorrer, assim como é muito habitual existirem várias versões do mesmo documento cujas alterações são mínimas porque as pessoas tendem a armazenar dados infinitamente, fazem uma pequena alteração e alteram logo o nome do documento, mas nunca chegam a apagar as versões mais antigas, mesmo depois de produzir a versão final gostam de permanecer com o histórico.

A gestão de informação e análise de dados, embora já seja o presente claramente que será e continuará a ser o futuro, procurando sempre evoluir e melhorar, não só os algoritmos como os próprios modelos a utilizar e a sua integração com a inteligência artificial, mas sem dúvida que o futuro se cruza e é compatível com o que foi feito com esta primeira abordagem. Em extremo máximo, na questão de partilha, podemos chegar ao ponto de ter alertas atempadamente de determinadas situações de acordo com determinados tipos de comportamento.

## CTEN STP RES Paulo Jorge Baptista das Neves Data: 2018-09-25

**(DITIC - Chefe do Núcleo de Resposta a Incidentes de Segurança da Informação)**

O protótipo executa o que propôs fazer, revelando a capacidade de encontrar dados considerados pessoais de acordo com as definições constantes no regulamento.

A grande vantagem que identifico é a possibilidade de fazer procuras de dados em ficheiros de estruturas distintas, incluindo os formatos mais comuns e de maior utilização pela organização.

Como melhoria, eventualmente seria interessante que após a identificação do número de ocorrências, fosse também possível apresentar graficamente uma correspondência dessas ocorrências com diferentes categorias.

Julgo que será uma mais valia para qualquer organização, e claro especificamente, para a Marinha. Principalmente como acelerador para a deteção de dados pessoais. Como detetor de desvios, eventualmente, mas apenas se for possível fazer uma correspondência entre o número de ocorrências com a natureza do ficheiro (tipo), de modo a identificar quais os processos que mais expõem mais dados. Em relação a uma possível ajuda do grau de privacidade, considero já mais difícil, ou pelo menos não me parece evidente a relação. A constatação de um numero de ocorrências num ficheiro, não significa obrigatoriamente que estes estejam comprometidos.

Como apreciação global, o produto está muito interessante e naturalmente tem todo o potencial de ser ou vir a ser uma ferramenta útil, com a particularidade de poder ser otimizada para a realidade da Marinha.

# Apêndice F Apresentação resultados *Edoclink*

A sessão da apresentação dos resultados dos testes realizados no estudo de caso ao *Edoclink*, ocorreu no dia 9 de outubro de 2018 e foi estrutura por duas componentes: (1) Apresentação dos objetivos e metodologia usada; e (2) Apresentação dos resultados da aplicação do protótipo aos dados estruturados e não estruturados do sistema *Edoclink* na Link Consulting, tendo como participantes os seguintes elementos:

- João Manuel Martins Barreira (Administrador);
- Fernando Jorge Costa da Silva Faria (Gestor de Produto e-DocLink);
- Ana Paula dos Reis Inácio (Gestora de Produto e-DocLink);
- Joana Rita Barreto Pontes Lagoute (Coordenadora de Gestão de Projetos).
- Paulo Alexandre Guerreio Fernandes (Chefe de Equipa);

Embora tenha sido uma sessão aberta com alguma interação durante a apresentação dos resultados, apenas foi entrevistado o administrador João Barreira cujo resumo dos comentários será descrito seguidamente e dizem respeito à parte de descoberta dos dados pessoais que possam violar a privacidade dos indivíduos, em especial aos resultados obtidos. Não foi aprofundada a parte relativa à metodologia de classificação do nível, da densidade de risco nem à AIPD.

No geral, o trabalho realizado parece muito interessante e apresenta já algumas, boas, potencialidades praticas, contudo, carece de alguns melhoramentos, nomeadamente na questão da afinação dos falsos positivos. Este ponto é importante para se ganhar uma maior credibilidade dos modelos criados. Neste sentido, focarei apenas quatro breves comentários, dois que considero que podem ser aspetos a melhorar e dois que considero pontos positivos do trabalho.

**Comentário 1.** Para alguns tipos de dados a precisão e a cobertura são baixas, e da discussão havida tal parece resultar de não ser usada nenhuma técnica que permita desambiguar os tipos de dados, por exemplo (1) o termo “Chaves” pode ser do tipo localidade, do tipo morada – “Avenida Defensores de Chaves” ou do tipo nome – “António Chaves”; e (2) o termo “Silva” pode ser um dado que possa querer significar um nome (dado pessoal) ou “silva” como simplesmente um nome de planta).

A discussão havida em torno deste assunto levou a sugerir que alguma análise de contexto, ou análise de grupos de palavras permitiriam reduzir as ambiguidades e melhorar precisão e cobertura. Neste caso concreto o modo de descoberta “padrões específicos” foi usado apenas para o tipo de dados “morada”. Contudo a utilização deste modo, pode ser adaptado e estendido a outros tipos de dados para permitir melhorar a precisão e cobertura, mesmo que para isso se tenha o custo ao nível de desempenho (tempo de análise).

**Comentário 2.** Durante a discussão foi referido que não tinha sido feita a exclusão de termos não significativos (preposições, conjunções, …), o que de certo modo, também, pode ter contribuído para o baixo valor da precisão para alguns dos tipos de dados. Foi referido e sugerido o uso da exclusão de

termos potencialmente não significativos, mas parece que esta exclusão deverá ser refletida e terá de ser realizada com equilíbrio, determinado quando são efetivamente termos não significativos.

**Comentário 3.** O protótipo fornece um relatório com identificação das pastas dos ficheiros e das tabelas das bases de dados onde foram descobertos dados pessoais, o tipo de dados pessoais descobertos e o nível de densidade do risco. A forma como os resultados são apresentados é muito favorável e agrega toda a informação relevante para a execução de ações de proteção dos dados, permitindo fazer uma priorização dos aspetos mais importantes, facilitando o trabalho de diagnostico.

**Comentário 4.** É introduzido o conceito de Densidade de Risco, que me pareceu muito importante para qualificar repositórios. Seria interessante aprofundar a utilidade do conceito, e refinar os parâmetros usados.