

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



Solução para a extinção da pegada digital de um utilizador

José António Teixeira Martins

Mestrado em Segurança Informática

Trabalho de Projeto orientado por:
Prof. Doutor António Casimiro Ferreira da Costa

Resumo

O termo “pegada digital” pode ser definido como um rasto tecnológico de algo ou alguém. Este poderá ter um impacto tanto positivo como negativo na imagem do utilizador ou organização. Embora cada qual seja responsável pela gestão da sua imagem na *internet*, manter um registo do rasto *online* é uma tarefa tediosa e inviável.

Face esta problemática, foi criada uma solução que visa gerir a pegada digital através da procura automatizada de contas do utilizador em diversas plataformas *web*. Esta solução permite ao utilizador descobrir onde se encontra registado o seu endereço de *e-mail* e, caso assim opte, cancelar ou remover o seu registo da plataforma pretendida.

O processo de remoção da conta de utilizador tem por base instruções de remoção fornecidas ao utilizador, podendo as mesmas ser executadas através da interface gráfica do *website* ou aplicação o que resulta num processo que pode ser tão simples como o clique num botão. Em alternativa, o utilizador pode optar pelo envio de um *e-mail* formal de pedido de remoção de registo da conta apelando às políticas de privacidade do país onde são processados os dados da mesma.

A solução proposta e implementada neste trabalho, Untrace, caracteriza-se pelo facto de juntar os processos de pesquisa e remoção numa única solução, utilizando o *e-mail* do utilizador.

Os resultados da avaliação de desempenho revelam que a solução Untrace apresenta vantagens em comparação com uma solução concorrente, solução Mine, através da utilização de um processo de pesquisa não-intrusiva e disponibilização de métodos de remoção de conta via interface ou *e-mail* com opções de customização.

Palavras-chave: pegada digital, sensibilização digital, análise pegada, rasto *online*, trilha digital

Abstract

The concept of digital footprint can be described as “someone’s technological tracks”. These tracks may have a negative or positive impact on one’s online image. Although each person must be responsible for their own online reputation, managing and keeping a close track of all personal online activity is both a tedious and inviable task.

To ease the process, a solution was created with the goal of managing a user’s digital footprint. This is achieved through the analysis of the user’s e-mail registration in online accounts across multiple web platforms. The solution allows the end-user to gain a better perception of its online presence and, if desired, supports the process of deleting such accounts.

Deleting an account is based on detailed instructions tailored to each website that are given to the user through the solution’s interface, which can be as simple as a click of a button. Additionally, the user may opt to send a formal e-mail request considering the privacy laws of the country or region where the data is being processed.

The value of the proposed solution, Untrace, is in its capability of merging both search and account removal processes while only requiring the end-user to submit his or her email address.

The results of the performance evaluation reveal that Untrace is advantageous when compared to a competitor solution, the Mine solution, through the non-intrusive search process and availability of account removal methods via interface or email with customization options.

Keywords: digital footprint, digital awareness, *online* reputation, *online* tracking, footprint analysis

Conteúdo

Capítulo 1	Introdução.....	1
1.1	Contextualização	1
1.2	Motivação.....	1
1.3	Problema.....	2
1.4	Instituição de acolhimento.....	3
1.5	Objetivos	3
1.6	Organização do documento.....	3
Capítulo 2	Contexto e trabalho relacionado.....	5
2.1	Presença <i>online</i> de um indivíduo.....	5
2.1.1	Pegada digital ativa	6
2.1.2	Porquê gerir a pegada digital.....	6
2.1.3	Como determinar a pegada digital.....	7
2.1.4	Como gerir a pegada digital	7
2.1.5	A gestão proativa da pegada digital	8
2.2	Ferramentas existentes	9
2.2.1	Pesquisa de endereço de <i>e-mail</i>	9
2.2.2	Pesquisa de informação comprometida.....	12
2.2.3	Redução da pegada digital.....	13
2.2.4	Discussão.....	18
2.3	Trabalho relacionado.....	20
Capítulo 3	Avaliação da necessidade de ferramentas da pegada digital	23
3.1	Questionário de avaliação da necessidade.....	23
Capítulo 4	Desenho e implementação da solução.....	29
4.1	Desenho da solução.....	29
4.1.1	Análise de requisitos	29
4.1.2	Aplicação das leis de privacidade	30
4.1.3	Âmbito de aplicação da solução.....	30
4.1.4	Aplicação da solução.....	31
4.1.5	Esquematização de processos.....	31
4.2	Implementação da solução	37
4.2.1	Política de privacidade	38
4.2.2	Infraestrutura de suporte.....	38
4.2.3	Base de dados de conhecimento.....	41
4.2.4	Componente <i>web</i>	44
Capítulo 5	Avaliação da solução.....	53
5.1.1	Avaliação quantitativa.....	53
5.1.2	Avaliação qualitativa.....	54
5.1.3	Discussão.....	56
Capítulo 6	Conclusões e trabalho futuro.....	59
6.1	Conclusões	59
6.2	Trabalho futuro.....	59
Anexo I	Política de privacidade.....	61
Bibliografia		63

Lista de figuras

Figura 2.1 – Botão de cancelamento de subscrição.....	8
Figura 2.2 - Menu para a remoção da conta do serviço Instagram.....	8
Figura 2.3 - Interface da solução Skymem.....	9
Figura 2.4 - Interface da solução Mailshunt.....	10
Figura 2.5 - Interface da solução Hunter.io.....	10
Figura 2.6 - Interface da solução haveibeenpwned.com	12
Figura 2.7 - Interface da solução dehashed	12
Figura 2.8 - Homepage da aplicação web Mine	13
Figura 2.9 - Detalhes sobre uma parte da pegada digital (eBay).....	14
Figura 2.10 - Pesquisa por nome de utilizador	15
Figura 2.11 - <i>Homepage</i> do serviço Unroll Me.....	15
Figura 2.12 - Menu de pesquisa da solução Just Delete Me	16
Figura 2.13 - Guias para a remoção de diferentes tipos de conta de utilizador.....	17
Figura 2.14 - Menu de pesquisa da aplicação web Account Killer	17
Figura 2.15 - Exemplos de guias disponíveis na solução Account Killer	17
Figura 4.1 - Diagrama sobre a recolha e processamento de informação	31
Figura 4.2 - Diagrama sobre a gestão da sessão do utilizador.....	32
Figura 4.3 - Diagrama sobre a pesquisa de contas do utilizador	33
Figura 4.4 - Diagrama sobre a classificação da pegada digital	34
Figura 4.5 - Diagrama sobre a gestão da informação	36
Figura 4.6 - Diagrama sobre a gestão da base de dados de conhecimento.....	37
Figura 4.7 - Diagrama da infraestrutura	38
Figura 4.8 - Arquitetura tecnológica	41
Figura 4.9 - Arquitetura da base de dados.....	42
Figura 4.10 - Excerto da <i>Query SQL</i>	42
Figura 4.11 - <i>Query SQL</i> agendada	43
Figura 4.12 - Excerto do script de validação mensal	43
Figura 4.13 - <i>E-mail</i> exemplo de validação mensal	44
Figura 4.14 - Diagrama da componente web	44
Figura 4.15 - Untrace – <i>homepage</i>	46
Figura 4.16 - Pedido de pesquisa de contas.....	46
Figura 4.17 - Excerto de código referente ao pedido à API.....	47
Figura 4.18 - Excerto de código de validação de registo no website Pinterest	48
Figura 4.19 - Excerto de código de validação de registo no website Instagram	48
Figura 4.20 - Página web de resultados, o e-mail encontra-se omitido.....	48
Figura 4.21 – Exemplo de classificação da pegada digital.....	49
Figura 4.22 - Guia para a remoção da conta Pinterest via interface do utilizador.....	49
Figura 4.23 - Pedido de remoção via e-mail	50
Figura 4.24 - Proposta de pedido via e-mail	50
Figura 4.25 - Formulário de confirmação	50
Figura 4.26 - Excerto de código JavaScript	51

Lista de tabelas

Tabela 2.1 - Comparação de soluções de pesquisa de <i>e-mail</i>	18
Tabela 2.2 - Comparação de soluções de pesquisa de informação comprometida.....	19
Tabela 2.3 - Comparação de soluções de redução da presença <i>online</i>	19
Tabela 3.1 - Resultados sobre a faixa etária	23
Tabela 3.2 - Intervalo de utilização da internet, em anos.....	24
Tabela 3.3 - N° contas online, por utilizador.....	24
Tabela 3.4 - N° de contas sob registo	24
Tabela 3.5 – Resposta sobre as ferramentas HaveIBeenPwned e Unroll Me	25
Tabela 3.6 - Familiaridade com o tema "Pegada Digital"	25
Tabela 3.7 - Nível de necessidade de uma solução de redução da pegada digital.....	26
Tabela 3.8 - Tipos de conta mais relevantes	26
Tabela 4.1 - Pesos atribuídos por tipo de <i>website</i>	35
Tabela 4.2 - Níveis de classificação da pegada digital.....	35
Tabela 4.3 - Comparação de <i>frameworks web</i>	39
Tabela 4.4 - Websites suportados.....	47
Tabela 5.1 – Métricas recolhidas.....	53
Tabela 5.2 - Resultados da pesquisa Untrace sobre o e-mail pessoal	54
Tabela 5.3 - Avaliação dos resultados obtidos	54
Tabela 5.4 - Número de contas canceladas ou em processo	55
Tabela 5.5 - Classificação de usabilidade da Untrace	55
Tabela 5.6 - Avaliação do nível de sensibilização para o tema "pegada digital"	55
Tabela 5.7 - Sugestões de melhoria.....	56
Tabela 5.8 - Comparação de funcionalidades entre Untrace e Mine.....	57

Capítulo 1

Introdução

1.1 Contextualização

A *internet* é como um universo digital, em constante expansão. A empresa CISCO, uma multinacional tecnológica reconhecida no mercado, prevê que até 2023 a percentagem de utilizadores da *internet* no mundo atinja os 66% (5.3 mil milhões) [1].

Existem diferentes formas de marcar presença na *internet*: através de subscrições a *newsletters* ou do registo de contas de utilizador, por exemplo em fóruns da *internet*, redes sociais ou *websites* de lojas *online*. Contudo, não existe uma etiqueta relativa à gestão desta presença pois nenhum dos utilizadores, ao aceder à *internet*, é informado das consequências positivas e negativas da mesma.

Aqueles que já fazem parte do mundo cibernauta há anos apercebem-se do quão facilmente se perde o controlo da própria presença *online*. Enquanto muitos poderão ver-se favorecidos pela visibilidade que esta oferece, outros poderão encontrar-se sob o outro lado da moeda onde a sua presença tem um impacto negativo na própria reputação. Este tipo de presença é também conhecido por pegada digital.

O trabalho de projeto que se segue aborda e apresenta uma solução tecnológica para a gestão da pegada digital incluindo métodos para a sua redução.

1.2 Motivação

Entende-se que não é fácil manter um registo de todos os locais na *internet* por onde circula a informação do utilizador pelo que em muitos dos casos o próprio nem consegue ter controlo sobre o processo. Uma das causas da perda deste controlo é dada pela forma como a informação do utilizador, seja o *e-mail*, número de telefone ou outros dados pessoais, pode ser exposta na *internet* fora de um âmbito inicialmente acordado nos termos e condições de um registo *online*. Esta exposição pode acontecer de forma intencional ou não-intencional.

A exposição intencional pode dar-se pela venda de informação para efeitos de *marketing* enquanto que na exposição não-intencional podem ocorrer incidentes de segurança relacionados com fuga de informação ou alterações à política de privacidade [2] que resultem na exposição de informação de uma forma não-controlada. Como consequência, utilizadores mal-intencionados poderão tomar partido da informação exposta e construir listas de *e-mails* posteriormente utilizadas em ataques de *phishing* massivos. Os ataques de *phishing* são retratados pelo envio de *e-mails* intencionalmente maliciosos que procuram levar o utilizador-alvo a clicar sobre um *link* e introduzir as suas credenciais ou outra informação confidencial. Exemplos comuns deste ataque cibernético podem ser vistos semanalmente nas páginas da Autoridade Tributária e Aduaneira [3] e dos CTT – Correios de Portugal [4].

Deste modo, a motivação do trabalho é combater a perda de perceção do utilizador sobre a sua pegada digital e por conseguinte promover a gestão da privacidade do mesmo. É certo que o

utilizador não tem, em grande parte, qualquer controlo sobre a informação que se encontra nos sistemas responsáveis pela mesma. No entanto, através do aumento da percepção e sensibilidade sobre a pegada digital será possível identificar os locais onde existem contas de utilizador e tomar uma ação sobre as mesmas reduzindo, tanto quanto possível, a exposição do utilizador.

Felizmente este não é um tema novo para a comunidade cibernauta pelo que atualmente existem ferramentas tecnológicas que procuram atingir este objetivo. Porém, estas são tendencialmente intrusivas e, como descrito no capítulo 3, colocam em causa a privacidade dos dados pessoais dos seus utilizadores. Assim sendo, a outra motivação deste projeto é conseguir alcançar os mesmos resultados sem recorrer a qualquer tipo de intrusão ou recolha de dados pessoais.

A Willway, instituição acolhedora deste projeto, tem como motivação a utilização desta solução como complemento às ferramentas de simulação de *phishing* e de auditoria de segurança, paralelamente aumentando a sensibilidade dos seus clientes para os diferentes riscos resultantes de uma pegada digital não controlada.

1.3 Problema

O problema a que este trabalho de projeto procura dar resposta envolve a gestão da pegada digital e pode ser encontrado nas questões sobre “porquê” e “como” efetivamente gerir uma pegada digital.

Relativamente à primeira, podemos primeiro observar que não há um consenso sobre a forma como um utilizador gere a sua presença *online* [5] [6] [7] sendo que daí surge a questão “Porque me devo preocupar com a minha pegada digital?”. Note-se ainda que, um utilizador desapegado do mundo digital ou que sinta que não tenha nada a ocultar na sua presença na *internet* poderia defender que a sua privacidade na *internet* não teria qualquer impacto na vida real. Todavia, este argumento não considera uma consequência da disseminação de informação por forma digital: é difícil controlar o que é dito sobre o utilizador, a que sítios *web* o nosso *e-mail* se pode encontrar associado ou por onde é que a nossa informação pessoal circula.

Num cenário fora do mundo digital, se ocorrer uma situação que afete a reputação de um indivíduo, este terá por norma meios físicos para tomar uma ação sobre a mesma. No caso da *internet* não é certo que tal seja possível ou que existam sequer pontos de contacto. Uma fotografia pode ser partilhada anonimamente e ser vista por quem menos esperamos como por exemplo o nosso futuro empregador, ou num caso oposto, a mesma poderá também ser utilizada a favor da reputação de um indivíduo [6]. Assim sendo, uma forma de controlar a divulgação não-solicitada da informação passa por uma gestão cuidada da pegada digital de um utilizador.

Em resposta à questão “Como posso identificar e reduzir a minha pegada digital?” encontramos o maior desafio da solução, o espaço de pesquisa sobre uma pegada digital. Sendo uma das principais formas de comunicação na *internet*, é também através do endereço de *e-mail* que se efetua o registo de contas *online*, todavia com a expansão da *internet* surge a problemática do acompanhamento: como validar o maior número de *websites* de forma a expandir ao máximo o espaço de pesquisa sobre a pegada digital? Na solução arquitetada, a resposta a este problema encontra-se na definição de um âmbito inicial e a sua expansão aquando do crescimento da solução. Por fim, sobre a redução da pegada digital surge o problema de como apagar um registo ou conta *online* quando em alguns casos [8] [9] [10] o processo não é intuitivo para o utilizador.

1.4 Instituição de acolhimento

A *Willway* é uma consultora especializada nas áreas de tecnologia e segurança de informação. Esta disponibiliza serviços de auditoria de segurança em diferentes vertentes dentro da área da ciber segurança. Os seus projetos incluem: análises de vulnerabilidades a aplicações *web*, aplicações *thick-client*, aplicações móveis, sensibilização de utilizadores sobre as temáticas do *phishing*, *ransomware* e *smishing* mediante o produto *Cybersecurity Cloud* e integração de soluções tecnológicas de segurança.

Em todos os projetos de auditoria existe uma fase inicial de recolha de informação. A título de exemplo, quando são analisados servidores ou aplicações *web* a recolha de informação segue um processo pré-definido [11]. No entanto, projetos de sensibilização de *spear-phishing* requerem a pesquisa sobre um ou mais utilizadores. A solução proposta neste trabalho visa auxiliar este último tipo de projetos e providenciar a futuros clientes um serviço de gestão da sua pegada digital, com foco na sensibilização sobre o tema.

1.5 Objetivos

Existem dois tipos de objetivos para este trabalho: os objetivos diretos (aqueles que serão avaliados) e os indiretos (aqueles que não sendo mensuráveis pressupõe-se que podem ser cumpridos). Entre os primeiros encontram-se a criação de um processo para a deteção de contas de utilizador *online*, a criação de um processo subsequente para remoção das mesmas via uma plataforma *web* de fácil utilização, a demonstração do valor desta solução comparativamente às existentes atualmente e a utilização da solução como complemento às atividades de auditoria e sensibilização para o *phishing* enquanto serviço prestado pela *Willway*.

Entre os objetivos indiretos encontram-se a preservação da privacidade do utilizador através da utilização de métodos de pesquisa da pegada digital não-intrusivos, o aumento da sensibilização do utilizador para o tema da pegada digital e a contribuição para o aumento da segurança das organizações utilizadoras da solução através da prevenção de uma pegada digital descuidada dos seus colaboradores.

1.6 Organização do documento

Abaixo encontra-se de forma estruturada a organização do documento:

- O capítulo 2 inicia o documento agora apresentado com o conceito de pegada digital. Posteriormente serão analisados os problemas identificados seguindo-se uma análise de ferramentas existentes que procurem dar resposta aos mesmos. O capítulo 2 termina com uma análise a outros trabalhos publicados na *internet* que envolvam uma discussão direta ou indireta sobre a presença *online*.
- No capítulo 3 é avaliada a perceção geral sobre o problema identificado através de um questionário de avaliação de necessidade. Este procura entender por amostragem o perfil do utilizador, qual o nível de perceção dos utilizadores face o tópico “pegada digital”, o nível de perceção sobre a dimensão da sua pegada digital e ainda quais os principais tipos de serviços que os mesmos preferem ver removidos, sejam redes sociais, fóruns, lojas *online* ou outros. A recolha e análise desta informação dita o caminho a seguir no desenho e implementação da solução, descritos no capítulo 4.

- O capítulo 4 inicia com um subcapítulo referente ao levantamento dos requisitos funcionais e não-funcionais da solução de forma a corresponder às necessidades anteriormente identificadas. Segue-se a análise sobre o âmbito e modo de aplicação da solução considerando o impacto de diferentes leis de privacidade. No subcapítulo de implementação será referida toda a base tecnológica da mesma, parte do código-fonte produzido, a forma como a solução interage com o utilizador-final e como o mesmo poderá utilizar a solução.
- No capítulo 5 discutir-se-á a avaliação da solução. A avaliação será feita através um projeto de implementação da solução em contexto profissional analisando os resultados obtidos e comparando-os face a utilização de uma ferramenta concorrente.
- No capítulo 6, serão referidos os próximos passos da solução e reflexões sobre eventuais lições aprendidas e limitações encontradas.

Capítulo 2

Contexto e trabalho relacionado

Entender o conceito de pegada digital é essencial para a percepção das consequências da falta de gestão da mesma. Uma pegada digital existe mesmo quando o utilizador tem pouca ou nenhuma presença *online* pois a crescente informatização dos processos administrativos, como por exemplo a marcação de consultas hospitalares, poderá requerer informação digital acerca do utente.

Neste capítulo será descrito o conceito de pegada digital e quais os tipos de pegada existentes. O foco da discussão irá incidir sobre a pegada digital ativa, sendo esta a que se encontra ao alcance dos utilizadores. O objetivo será explicar e justificar o processo de gestão da pegada apresentando ferramentas que visam auxiliar nessa mesma gestão.

A análise das ferramentas é seguida de uma discussão sobre as mesmas com o objetivo de entender quais as funcionalidades mais importantes numa solução deste tipo.

De forma a complementar este estudo segue-se ainda um subcapítulo dedicado à análise de trabalhos publicados sobre a pegada digital que ajudam na compreensão do tema.

2.1 Presença *online* de um indivíduo

Uma pegada é definida como um conjunto de marcas deixadas pela passagem de algo ou alguém [12]. Por sua vez, num contexto informático, “digital” [13] é definido por tudo aquilo que envolve tecnologia eletrónica, sejam por exemplo computadores ou a *internet*. Através da junção destes dois conceitos podemos então definir que a “pegada digital” se trata de um rasto tecnológico de algo ou alguém.

Girardin et al [14] subdividem este rasto em dois: a pegada digital passiva e ativa. A primeira refere-se ao tipo de informação recolhida de forma automática pelos sistemas de informação seja por questões de segurança da informação ou avaliação de desempenho e um claro exemplo disto é dado pelo processo de telemetria [15]. Este processo foi inicialmente utilizado para a monitorização remota do estado de dispositivos através da recolha de informação disponibilizada por sensores colocados no mesmo. Na transição da telemetria para o mundo digital atual podemos observar que o Windows implementa este processo com o objetivo de melhorar a qualidade e segurança do seu serviço. Contudo, os dados recolhidos pelo sistema operativo Windows e enviados de volta para a Microsoft podem chegar a incluir o texto introduzido pelo utilizador nos últimos 30 minutos [16]. Outro tipo de recolha da pegada digital de forma passiva acontece sob a forma de alteração da política de privacidade de uma aplicação. Como exemplo disto temos o caso de estudo da Audacity, um *software* de edição de som que, em junho de 2021, alterou a sua política de privacidade para incluir a recolha de informação acerca do sistema operativo dos seus utilizadores [17]. Infelizmente a gestão deste tipo de pegada digital requer ainda processos manuais de revisão de políticas e de comportamento aplicacional, tarefas pouco automatizáveis ao dia de hoje.

A pegada digital ativa, objeto de estudo deste trabalho, refere-se ao tipo de rasto criado voluntariamente por um indivíduo ou organização. Como exemplos deste tipo de pegada temos a

criação de contas de utilizador, a subscrição a *newsletters*, o preenchimento de formulários *online*, a aceitação da utilização dos *cookies* ou a utilização da localização geográfica em serviços *online*. À partida, o utilizador consegue ver e é parte ativa da criação deste tipo de pegada digital. Deste modo, a principal diferença entre a pegada digital ativa e passiva é dada pelo nível de visibilidade que o utilizador consegue ter sobre a mesma.

2.1.1 Pegada digital ativa

Ainda que a pegada digital ativa esteja associada a um comportamento voluntário do indivíduo na *internet*, existe nesta uma subcomponente passiva [14]. Quando o utilizador se regista numa rede social este interage de forma passiva por meio das ações das suas ligações na rede [18]. O utilizador não tem controlo sobre o que os seus “amigos”, ligações ou conhecidos possam publicar sobre o mesmo. Deste modo, a informação da pegada digital ativa inclui também fotografias, comentários, “gostos” e outro tipo de identificações.

Para estes casos a gestão da pegada digital torna-se num processo novamente manual pois as definições de privacidade das redes sociais [19] não impedem o comportamento descuidado de outros utilizadores. Sendo que este tipo de informação é uma componente da pegada digital ativa, as formas de mitigação desta passam muitas vezes pela remoção ou desativação da conta de utilizador.

Adicionalmente, utilizadores ou organizações externas podem tomar proveito desta exposição para fins lucrativos como a venda de informação ou fins mal-intencionados como os ataques *spear-phishing*, um cenário em que um ataque de *phishing* é feito à medida de interesses do utilizador-alvo.

2.1.2 Porquê gerir a pegada digital

Existe algum trabalho prévio que evidencia que a pegada digital tem impacto na componente social e profissional de um indivíduo [6] [7] [20]. Ao partilharmos informação *online* estamos a construir uma *persona*, uma identidade que pode ser vista por qualquer pessoa com acesso à *internet* incluindo amigos, familiares e/ou recrutadores de emprego [21]. Neste último caso, dependendo da informação encontrada, a pegada digital poderá ter um impacto positivo ou negativo no processo de recrutamento. Sobre o impacto na componente social encontra-se o exemplo referido anteriormente: o que os demais partilham acerca do indivíduo sem autorização prévia poderá também afetar a reputação social do mesmo. Deste modo o utilizador deverá procurar cuidar da sua presença *online*.

Contudo, a ausência de uma presença *online* também pode produzir efeitos indesejados. Profissionalmente, candidatos com maior visibilidade na *internet*, desde que positiva, terão uma vantagem face os restantes concorrentes. Esta presença positiva inclui, por exemplo, a criação de um *portfolio* pessoal, contribuições para a comunidade ou a participação em atividades de voluntariado. Assim, haverá uma propensão para continuar o processo de recrutamento com o candidato cujo nome se encontre associado a publicações ou eventos de boas-causas quando comparado com outro candidato cuja pesquisa não retorne qualquer tipo de informação pertinente.

Na ótica do relacionamento social, a ausência de um indivíduo *online* pode ser vista de ambas perspetivas, o que envolve uma discussão de opinião pessoal pouco relevante para o restante trabalho.

Existe, portanto, uma certa discordância entre os utilizadores sobre se a divulgação de informação *online* se apresenta como um benefício ou uma preocupação [5] visto que o contexto

em que é feita pode ter variados resultados. Assim sendo, a privacidade do utilizador pode beneficiar de uma gestão da sua pegada digital.

2.1.3 Como determinar a pegada digital

Um utilizador poderá determinar a sua pegada digital de duas formas, pessoal ou interpessoal.

O primeiro método envolve a análise manual ou automatizada da informação pessoal do próprio. O utilizador pode pesquisar manualmente a sua caixa de correio em busca de mensagens de *e-mail* relativas a registos, verificar a que *websites* estão associados os *cookies* do seu navegador de *internet*, verificar *passwords* guardadas e a que serviços pertencem, verificar SMSs, registos de chamadas telefónicas e até mesmo cartões de fidelização onde possa ter referido o seu endereço de *e-mail*. No entanto, a utilização de ferramentas para a automatização esta pesquisa pode requerer que o utilizador forneça acessos à sua caixa de correio ou ao próprio sistema operativo. Este nível de intrusão coloca em risco a privacidade do utilizador pois que tal ferramenta fosse de domínio publico, a maioria dos utilizadores teria de confiar quase cegamente na correta gestão e proteção do acesso aos seus dados.

Os métodos de pesquisa interpessoal visam utilizar apenas recursos disponíveis publicamente minimizando a interação com o utilizador. Um método utilizado na determinação da pegada digital ativa é através do conceito de OSINT – *Open Source Intelligence*.

OSINT – *Open Source Intelligence*

Este conceito refere-se à pesquisa de informação através do uso de fontes de informação públicas. A *OSINT Framework* [22] contém um guia para o tipo de fontes que podem ser exploradas de forma a obter informação sobre algo ou alguém.

A *framework* disponibiliza um conjunto variado de temas a pesquisar, desde a pesquisa por nome de utilizador, fontes de partilha de informação sobre *malware* a fontes de formação sobre a própria OSINT. De forma a minimizar a informação necessária por parte do utilizador serão apenas utilizadas fontes de informação onde o *e-mail* é suficiente para uma pesquisa da pegada digital.

Dentro deste critério incluem-se as seguintes áreas da OSINT:

- Fontes de pesquisa de endereço de *e-mail*;
- Fontes de pesquisa de informação comprometida;
- Ferramentas relacionadas com a privacidade do utilizador e redução da presença *online*.

Por conseguinte, para cada uma destas áreas encontram-se disponíveis diferentes ferramentas que podem auxiliar no processo da gestão da pegada digital. A sua forma de utilização das mesmas e respetiva análise é descrita nos próximos capítulos.

2.1.4 Como gerir a pegada digital

Tendo definido as diferentes formas de determinar uma pegada digital resta ainda discutir como as colocar em prática. A solução proposta prevê que a pegada digital seja minimizada através da remoção do maior número de registos do utilizador na *internet*. Contudo, cabe ao mesmo, em alternativa à remoção, decidir se mantém algum registo *online* como parte de uma

gestão ponderada da sua pegada digital. A metodologia utilizada é aplicável tanto aos métodos pessoais como interpessoais referidos anteriormente.

Atualmente existem duas formas de um utilizador tentar remover uma conta, subscrição ou outro tipo de registo *online*:

1. Utilização da *interface* da plataforma;
2. Pedido de remoção formal.

No primeiro método estão incluídas as ações que requerem a ação do utilizador como por exemplo clicar nos *URLs* “Cancelar a sua subscrição” (Figura 2.1) ou aceder aos menus de remoção de conta de uma plataforma *web*, habitualmente disponíveis no painel de administração de conta (Figura 2.2).



Figura 2.1 – Botão de cancelamento de subscrição

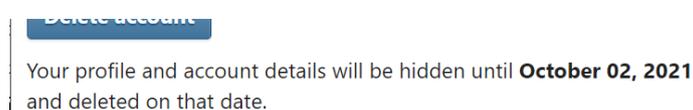


Figura 2.2 - Menu para a remoção da conta do serviço Instagram

A utilização do segundo método envolve o envio de um pedido formal, por norma via *e-mail*, para o serviço em questão. A título de exemplo, a remoção de uma conta registada na Amazon terá de ser feita via *e-mail* onde é pedido que o utilizador escreva uma breve mensagem sobre a razão desta remoção.

Apesar de a Amazon não possuir requisitos exigentes quanto à estrutura de um *e-mail* formal fará sentido a criação de um modelo de *e-mail* que possa ser utilizado transversalmente nos diversos serviços na *internet*. De forma a sustentar a legitimidade deste pedido também será útil envolver as leis de privacidade ao alcance do utilizador. Por exemplo, dentro da União Europeia um *e-mail* para a remoção de conta terá um efeito maior se apelar ao RGPD (Regulamento Geral sobre a Proteção de Dados) do que se simplesmente contiver um texto cordial. As empresas sediadas na União Europeia estão sob o abrigo deste regulamento que visa proteger a privacidade dos utilizadores.

Entre as diferentes medidas impostas encontra-se o artigo 17º, intitulado “direito ao esquecimento” que permite ao utilizador requisitar a remoção de toda a informação pessoal presente nos sistemas da organização a quem se dirige. As diferentes regiões onde esta metodologia pode ser aplicada e as respetivas leis de privacidade serão mencionadas e discutidas durante a fase do desenho da solução.

2.1.5 A gestão proativa da pegada digital

Durante o processo de redução e gestão da pegada digital o utilizador-alvo poderá deparar-se com um número volumoso de contas de utilizador. Dado que os processos formais poderão durar semanas, fará sentido utilizar a solução de uma forma espaçada. Recomenda-se uma utilização mensal, trimestral ou semestral de acordo com a frequência habitualmente baixa com que um utilizador cria novas contas o que permitirá ao utilizador entender se existem novos registos.

Nos casos em que surjam novos registos inesperados o utilizador poderá ainda concluir qual das plataformas-alvo anteriormente analisadas possivelmente partilhou a sua informação. Esta gestão e utilização proativa da solução traz, portanto, vantagens ao nível da otimização da pegada digital bem como a possibilidade de descoberta de violações de políticas de privacidade.

Como forma de motivar a reutilização da plataforma será disponibilizado ao utilizador um nível de avaliação da sua pegada digital baseado no rácio do utilizador entre contas registadas e o total de *websites* disponíveis para pesquisa.

2.2 Ferramentas existentes

De seguida serão descritas ferramentas disponíveis para a análise da pegada digital ativa de um utilizador. Esta análise envolve a avaliação sobre três tópicos: interface, funcionalidades disponíveis e funcionalidades adicionais.

Na avaliação da interface será medida a facilidade com que a interface da solução cumpre o seu propósito. Na avaliação das funcionalidades disponíveis serão avaliadas as funcionalidades disponibilizadas pela ferramenta que correspondam ao objetivo inicial (redução da pegada digital ativa). Na avaliação de funcionalidades adicionais será mais vantajosa aquela que apresentar utilidades fora do âmbito necessário à redução da pegada digital (sistema de autoavaliação, pontuação e/ou outras).

2.2.1 Pesquisa de endereço de *e-mail*

A utilidade deste tipo de ferramentas serve o objetivo da Willway para os casos em que a mesma não tem nenhum conhecimento de endereço de *e-mail*. Durante o processo auditoria de segurança é comum não haver qualquer tipo de informação acerca dos *e-mails* da empresa sob auditoria, esta pesquisa é parte integrante do processo. Para testar e avaliar as seguintes ferramentas foi utilizado o endereço de *e-mail* “jmartins” e o domínio de *e-mail* “willway.pt”.

Skymem



Figura 2.3 - Interface da solução Skymem

Skymem [23] é um serviço *online* que dado um domínio de *e-mail* (Figura 2.3) efetua a pesquisa de contas do mesmo recolhidas através de fontes públicas na *internet*. Os utilizadores de outros serviços da Skymem também poderão contribuir para a população do seu repositório de *e-mails* através de uma participação voluntária.

Interface

A solução pode ser utilizada consultando apenas a página principal efetuando a pesquisa pretendida. Os resultados são obtidos de forma rápida.

Funcionalidades

Pesquisa limitada de *e-mails* associados ao domínio. O acesso aos restantes resultados está dependente da subscrição paga da solução.

Outras funcionalidades

Nenhuma encontrada.

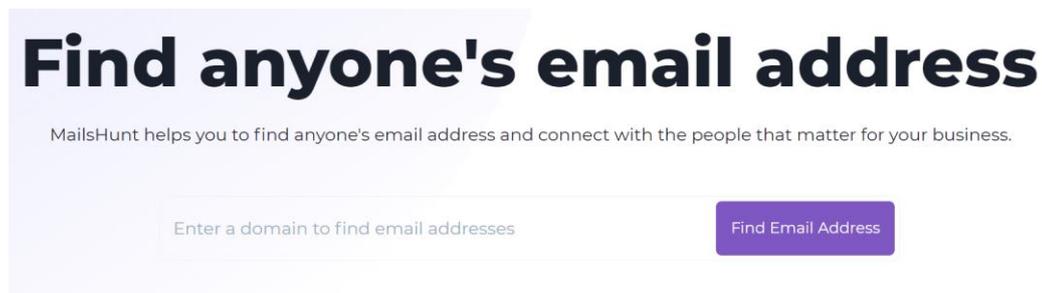
Mailshunt

Figura 2.4 - Interface da solução Mailshunt

Mailshunt [24] especializa-se na extração de *e-mails* relacionados com um determinado *website*. Este serviço apresenta uma oferta de validação de *e-mails* por forma a garantir que pertencem a utilizadores reais (Figura 2.4).

Interface

A solução pode ser utilizada consultando apenas a página principal efetuando a pesquisa pretendida.

Funcionalidades

Pesquisa limitada de *e-mails* associados ao domínio. Não existe indicação plausível de que foram detetados mais resultados válidos, no entanto a solução sugere a subscrição paga do serviço.

Outras funcionalidades

Nenhuma encontrada.

Hunter.io

Figura 2.5 - Interface da solução Hunter.io

Este serviço [25] destaca-se pelo número de resultados da pesquisa de *e-mails* (Figura 2.5) graças à sua volumosa base de dados [26]. É possível aceder a um resultado limitado por 25 *e-mails* gratuitamente.

Interface

A aplicação é de fácil utilização através de uma única página *web*. A aplicação despoleta periodicamente um aviso publicitário sobre a sua extensão disponível para o navegador de *internet*.

Funcionalidades

Pesquisa limitada de *e-mails* associados ao domínio. Opção da criação de uma conta de avaliação com direito a 25 resultados por pesquisa.

Outras funcionalidades

É devolvida informação sobre as fontes nas quais foram encontradas as referências aos *e-mails*.

Infoga

Infoga [27] é uma solução de código-livre desenvolvida em Python que utiliza motores de busca e consulta bases de dados de *e-mails* comprometidos no processo de pesquisa.

Interface

Não existente, trata-se de uma solução que é utilizada via terminal do sistema operativo do utilizador.

Funcionalidades

Pesquisa sem limites através de um conjunto de motores de busca.

Outras funcionalidades

Disponibilização do código-fonte da solução. Informação sobre as fontes de origem dos resultados.

theHarvester

theHarvester [28] é uma solução de código-livre desenvolvida em Python que utiliza motores de busca, consulta bases de dados de *e-mails* comprometidos, registos DNS, ferramentas de detecção de vulnerabilidades entre outras fontes para a pesquisa de *e-mails*.

Interface

Não existente, trata-se de uma solução que é utilizada via terminal do sistema operativo do utilizador.

Funcionalidades

Pesquisa sem limites através de um conjunto de motores de busca e serviços *web*. A utilização dos serviços *web* requer um registo prévio nos mesmos.

Outras funcionalidades

Disponibilização do código-fonte da solução. Informação sobre as fontes de origem dos resultados. Elevado número de fontes de pesquisa. Possibilidade de recolha de informação técnica adicional como nomes, subdomínios e endereços IP.

2.2.2 Pesquisa de informação comprometida

Este tipo de ferramentas auxilia na determinação da pegada digital com a benesse de informar o utilizador de que as suas credenciais poderão ter sido comprometidas.

Através dos resultados da pesquisa sobre serviços de informação comprometida é possível apurar algumas das plataformas *web* nas quais o endereço de *e-mail* do utilizador se encontra registado, complementando o universo de pesquisa da pegada digital. Para testar e avaliar as seguintes ferramentas foi utilizado o endereço de *e-mail* “jmartins” e o domínio de *e-mail* “willway.pt”.

Have I been pwned

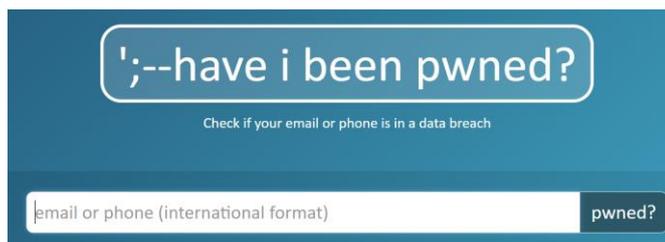


Figura 2.6 - Interface da solução haveibeenpwned.com

A solução HaveIBeenPwned [29] tem como objetivo analisar e informar o público em geral acerca de comprometimentos de credenciais detetados na *internet*. Estes incidentes de segurança contêm, por norma, listas de *e-mails* provenientes de fugas de dados.

Interface

A aplicação é fácil de utilizar bastando inserir o *e-mail* a pesquisar na página principal (Figura 2.6).

Funcionalidades

Pesquisa por *e-mail* sobre uma base de dados de fugas de informação em constante atualização.

Outras funcionalidades

Pesquisa por domínio disponível mediante devida autorização para tal. Pesquisa por *passwords* comprometidas.

Dehashed

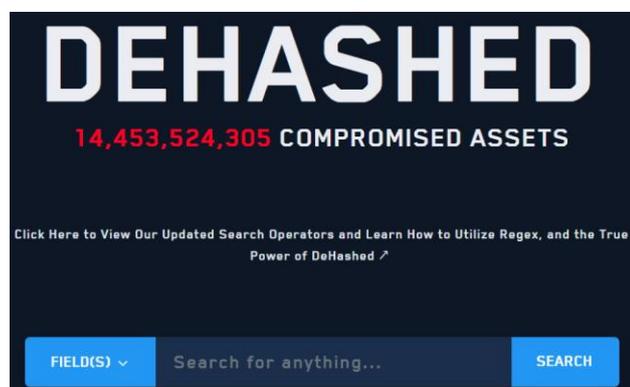


Figura 2.7 - Interface da solução dehashed

O serviço Dehashed [30] comporta-se de forma semelhante a HaveIBeenPwned, no entanto disponibiliza um acesso às fontes de informação recolhidas através de uma subscrição paga.

Interface

A aplicação requer um registo prévio, não é necessária a validação do *e-mail*.

Funcionalidades

Pesquisa por *e-mail* sobre uma base de dados de fugas de informação. Determinação da pegada digital através dos resultados obtidos na pesquisa. É necessária uma subscrição paga para aceder aos resultados concretos da pesquisa.

Outras funcionalidades

Diferentes tipos de pesquisa possível (Figura 2.7), por domínio, nome de utilizador, endereço *IP*, nome, morada e contacto telefónico.

2.2.3 Redução da pegada digital

As ferramentas utilizadas na redução da presença *online* identificadas utilizam ambos métodos intrusivos e não intrusivos nos seus processos de deteção. Como exemplo do primeiro temos as soluções Mine e Unroll Me que através do acesso à caixa de *e-mails* do utilizador efetuam pesquisas sobre os mesmos por forma a identificar plataformas *web* ou serviços no qual o *e-mail* se encontre registado. Não havendo uma visibilidade sobre o código-fonte destas soluções não é possível determinar a forma exata de como as mesmas utilizarão a permissão concedida sobre a leitura de *e-mails* do utilizador final, ainda que a forma de utilização dos dados esteja descrita nas respetivas políticas de privacidade [31] [32].

Em contrapartida, as soluções Usersearch e Holehe executam o mesmo processo de deteção sem que seja necessário o acesso a dados do utilizador. Adicionalmente foram identificadas soluções que apesar de não conterem nenhum processo de deteção de contas de utilizador, possuem informação sobre como remover os registos em diversas plataformas *web*. Para efeitos de teste foram utilizadas diferentes contas de *e-mail* pessoal subscritas a serviços conhecidos.

Mine



Figura 2.8 - Homepage da aplicação web Mine

A Mine [33] é uma plataforma *online* que permite ao utilizador reduzir a sua pegada digital através do envio de pedidos formais de remoção de dados pessoais.

Interface

No primeiro acesso à aplicação *web* (Figura 2.8) esta solução possui apenas uma página *web* com uma descrição muito sucinta sobre o serviço que oferecem. Nesta descrição encontram-se atalhos para a respetiva política de privacidade, termos e condições e um consentimento sobre o uso de *APIs* (*Application Programming Interfaces*).

Após o registo da conta de utilizador é fácil verificar que a aplicação é intuitiva. Durante a pesquisa da pegada digital são apresentadas curtas mensagens de *feedback* úteis para manter atenção do utilizador durante um processo que é naturalmente demorado.

Para cada um dos serviços encontrados na pegada digital do utilizador existe um botão de “reclaim” ou de validação de detalhes sobre o serviço (Figura 2.9). Na visão detalhada do mesmo é ainda possível aceder a uma avaliação de risco do serviço baseada no histórico de violação de dados do serviço. A fórmula utilizada neste cálculo não é mencionada.

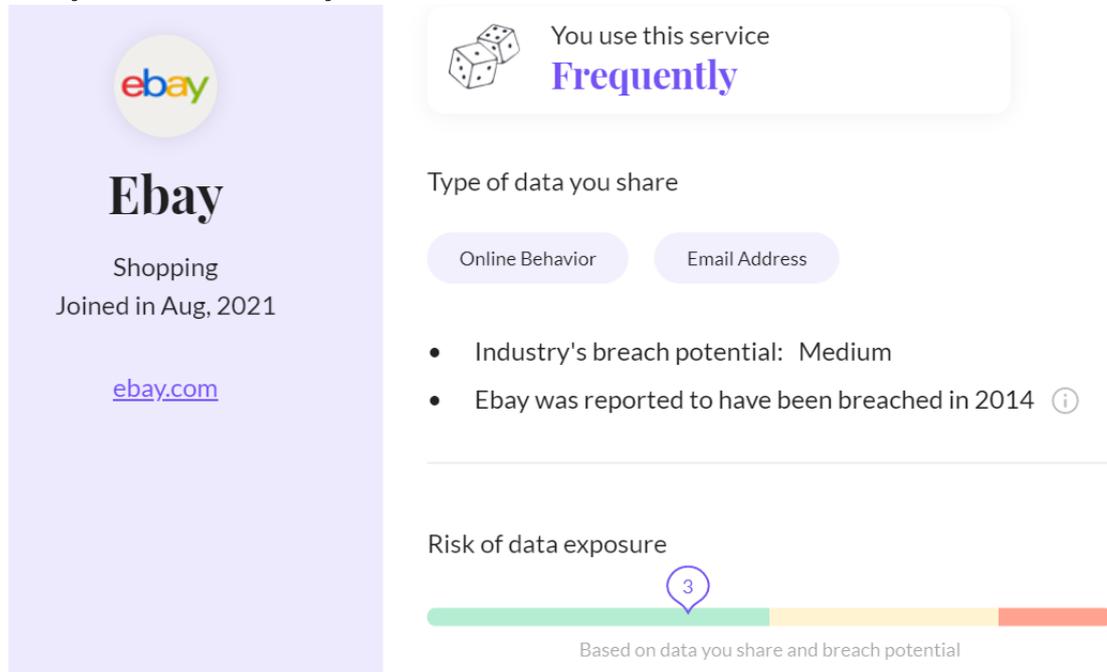


Figura 2.9 - Detalhes sobre uma parte da pegada digital (eBay)

Funcionalidades

Existem apenas duas funcionalidades disponíveis: remover os dados pessoais ou ver detalhes sobre o serviço. O processo de remoção é sempre feito através do envio de um modelo de *e-mail* formal. A estrutura do mesmo nunca muda e não utiliza nenhuma referência a leis de privacidade.

A utilização do serviço está dependente da cedência de um acesso à caixa de correio do utilizador.

Outras funcionalidades

Para cada um dos serviços detetados existe uma escala de exposição. Encontra-se sempre disponível um menu com perguntas frequentes e ajuda.

A interface mostra mensagens enquanto se espera pelos resultados, um processo que é de esperar ser demorado. A leitura destas mensagens requer permissões de leitura da caixa de entrada do *e-mail*.

Usersearch.org

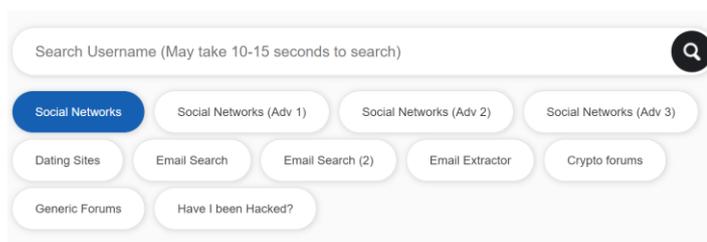


Figura 2.10 - Pesquisa por nome de utilizador

O serviço Usersearch [34] permite ao utilizador pesquisar um nome utilizador sob diferentes tipos de plataformas (Figura 2.10). As diferentes formas de pesquisa permitem ao utilizador detetar registos em redes sociais, fóruns e na base de dados HaveIBeenPwned.

Interface

A interface desta aplicação *web* é composta por uma única página *web*. No formulário de pesquisa é possível filtrar a procura. Existe uma descrição bastante curta para cada um destes filtros.

Funcionalidades

Existe uma funcionalidade de pesquisa via nome de utilizador. Para efetuar a pesquisa via *e-mail* será necessário subscrever o serviço pago.

Outras funcionalidades

Não foi detetada nenhuma funcionalidade adicional.

Unroll Me

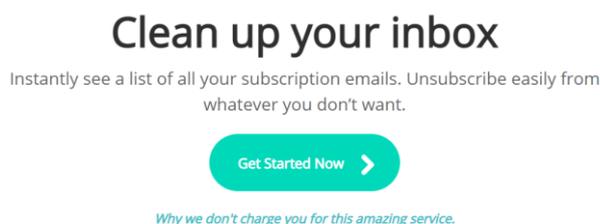


Figura 2.11 - Homepage do serviço Unroll Me

Unroll Me [35] permite ao utilizador detetar as suas subscrições de *e-mail* ativas e desativá-las, à semelhança de um clique em “Cancelar a sua subscrição” (Figura 2.11).

Interface

A interface desta solução assemelha-se bastante à da Mine. Trata-se de uma interface intuitiva que guia o utilizador durante todo o processo de deteção de subscrições a *newsletters*.

Funcionalidades

A plataforma Unroll Me permite ao utilizador remover subscrições a *newsletters* do seu *e-mail* pessoal. Para tal, o utilizador deverá conceder acesso à sua caixa de *e-mails*. O serviço trata então de analisar a caixa de entrada do utilizador e apresenta os serviços aos quais se encontra subscrito.

Esta solução encontra-se apenas disponível para residentes fora da União Europeia.

Outras funcionalidades

O serviço oferecido, apesar de limitado às *newsletters*, é gratuito.

Holehe

Holehe [36] é uma solução de código-livre desenvolvida em Python que utiliza as funcionalidades de registo e recuperação de *password* para detetar a presença de um registo de *e-mail*.

Interface

Não existente, trata-se de uma solução que é utilizada via terminal do sistema operativo do utilizador.

Funcionalidades

A ferramenta de código-livre Holele permite ao utilizador, através da introdução do seu *e-mail*, entender em que *websites* se encontra registado. Esta informação é obtida através do preenchimento dos formulários do tipo “Esqueci a minha palavra-passe”. Ao analisar a resposta obtida destes formulários a ferramenta consegue confirmar a existência do registo do *e-mail*. Esta metodologia de validação de registo é bastante útil à solução proposta, contudo, a ferramenta não refere como proceder à remoção das contas.

Outras funcionalidades

Não foram encontradas funcionalidades adicionais.

Just delete me



Figura 2.12 - Menu de pesquisa da solução Just Delete Me

Esta solução *web* [37] contém informação de processos de remoção de contas (Figura 2.12) em vários serviços *online*.

Interface

A interface da aplicação *web* é intuitiva. Comparando graficamente com as restantes, encontra-se longe de ser a mais apelativa.

Funcionalidades

A aplicação *web* permite ao utilizador aceder aos processos de remoção de contas de uma extensa coleção de aplicações e serviços. O objetivo dos autores é o de desmistificar o processo por vezes obscuro de suspensão dos serviços. Apesar da ferramenta não conter nenhum processo automático contém informação bastante valiosa para a solução proposta. A figura abaixo ilustra o tipo de informação disponível.

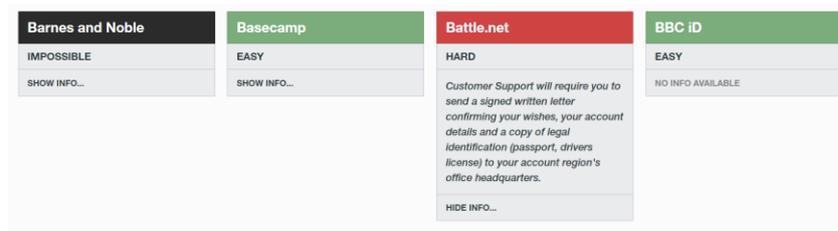


Figura 2.13 - Guias para a remoção de diferentes tipos de conta de utilizador

Existem atualmente 4 níveis de dificuldade (Figura 2.13) para a remoção de uma conta de utilizador entre “fácil” e “impossível”. Não existe uma explicação para o uso esta escala. No entanto, é possível observar que os serviços de fácil remoção podem ser removidos através da sua interface *web*. Por outro lado, nos serviços identificados como “impossível” a justificação para tal varia: uns serviços não conseguem remover a informação, outros simplesmente recusam-se a fazê-lo.

Outras funcionalidades

Não foram encontradas funcionalidades adicionais relevantes.

Account Killer



Figura 2.14 - Menu de pesquisa da aplicação web Account Killer

A solução *web* Account Killer [38] disponibiliza uma base de dados de guias para a remoção de contas *online* à semelhança da solução Just Delete Me.

Interface

A utilização desta interface (Figura 2.14) requer que o utilizador abra uma nova janela para cada consulta feita. Por esta razão, a usabilidade desta aplicação perde-se dada a extensa coleção de guias (Figura 2.15).



Figura 2.15 - Exemplos de guias disponíveis na solução Account Killer

Funcionalidades

Esta aplicação *web* é semelhante à Just Delete Me. É possível consultar guias de remoção de contas em diversos serviços e aplicações.

Cada um dos guias possui informação detalhada caso o utilizador necessite de entrar em contacto com o serviço em questão.

Outras funcionalidades

A solução descreve o propósito de cada um dos serviços *web* para os quais possui um guia.

DeleteMe

O serviço DeleteMe [39] especializa-se na remoção de informação pessoal presente no motor de busca Google. Trata-se de um serviço comercial pago direcionado a empresas ou indivíduos que procurem otimizar a sua reputação neste motor de busca.

Interface

Dada a vertente comercial deste serviço não foi possível validar a interface do mesmo.

Funcionalidades

Esta aplicação *web* auxilia na remoção da informação pessoal presente em pesquisas do Google.

Outras funcionalidades

A vertente comercial desta solução não permitiu a identificação de funcionalidades adicionais.

2.2.4 Discussão

As tabelas seguintes sumarizam as vantagens e desvantagens das ferramentas anteriormente apresentadas.

Soluções referentes à pesquisa de endereço de *e-mail*

Tabela 2.1 - Comparação de soluções de pesquisa de *e-mail*

Solução existente	Aspetos positivos	Aspetos negativos
Skymem	Execução rápida da pesquisa.	Limitação comercial de 5 resultados.
Mailshunt	Execução rápida da pesquisa.	Limitação comercial sobre os resultados; Resultados obtidos não aparentam ser fiáveis.
Hunter.io	Pesquisa rápida; Informação sobre a fonte dos resultados.	Solução comercial limitada a 25 resultados na versão gratuita; Pop-up ocasional de publicidade sobre a extensão do navegador.
Infoga	Conjunto de fontes limitado a motores de busca <i>web</i> .	Sem interface <i>web</i> .
theHarvester	Conjunto alargado de fontes de pesquisa, incluído diversos serviços <i>web</i> .	Sem interface <i>web</i> .

De entre as soluções com interface gráfica (Tabela 2.1) conclui-se que Hunter.io seria a melhor opção para corresponder à necessidade deste tipo de ferramentas. No entanto, a utilização da

solução sem interface gráfica theHarvester quando propriamente configurada permite ao utilizador obter um maior número de resultados que qualquer outra ferramenta proposta.

Soluções referentes à pesquisa de informação comprometida

Tabela 2.2 - Comparação de soluções de pesquisa de informação comprometida

Solução existente	Aspetos positivos	Aspetos negativos
Have I been Pwned	Bases de dados volumosa.	Não apresenta informação detalhada sobre os resultados.
Dehashed	Permite uma pesquisa mais detalhada através de diferentes campos.	Requer um registo prévio à pesquisa; Requer uma subscrição paga para a consulta detalhada dos resultados.

Existem poucas ferramentas que mantenham um registo atualizado de informação comprometida (Tabela 2.2). Entre as duas encontradas podemos observar que Dehashed oferece um maior detalhe sobre os resultados encontrados após a subscrição dos seus serviços. Para efeitos da determinação da pegada digital Have I been pwned é a solução ideal na conjugação com outras ferramentas.

Soluções referentes à redução da presença *online*

Tabela 2.3 - Comparação de soluções de redução da presença *online*

Solução existente	Aspetos positivos	Aspetos negativos
Usersearch.org	Fácil utilização.	Acesso <i>premium</i> ; Código-fonte privado.
Unroll Me	Modo de atuação transparente; Interface gráfica atrativa.	Intrusiva; Apenas remove <i>newsletters</i> .
Holehe	Versátil na integração com outras tecnologias.	Sem interface ou aplicação <i>web</i> .
Just delete me	Base de dados com informação relevante sobre a remoção de contas.	Não permite a pesquisa de utilizadores / <i>e-mails</i> .
Account Killer	Base de dados com informação relevante sobre a remoção de contas.	Não permite a pesquisa de utilizadores / <i>e-mails</i> .
DeleteMe	Nenhum encontrado.	Solução comercial; Restrita a um conjunto limitado de <i>websites</i> .
Mine	Cumprir com o objetivo da solução proposta.	Falta de informação sobre como remover uma conta via interface; Utilização de um modelo de <i>e-mail</i> com oportunidades de melhoria; Intrusiva.

De entre as soluções anteriores (Tabela 2.3) podemos observar que as mais úteis ao projeto serão a Holele, devido à sua integrabilidade com o resto da solução proposta e a Mine, pois é a única ferramenta que incorpora as duas fases de pesquisa e controlo da pegada digital. Na outra ponta do espectro encontram-se as soluções Unroll Me, limitada pela capacidade de remoção de apenas subscrições de *e-mails* e ainda DeleteMe, uma solução comercial que abrange apenas a informação presente nas pesquisas do Google.

Apesar das suas diferenças na utilidade para o desenvolvimento da solução, todas poderão contribuir, no mínimo, como fontes de informação para a base de dados de conhecimento. As soluções Just Delete Me, Account Killer e Have I Been Pwned são as mais adequadas pois dedicam-se à disponibilização de informação.

A base de dados de conhecimento da solução proposta tem como objetivo recolher toda a informação acerca dos métodos formais ou informais de remoção de uma conta de utilizador bem como o processo original de verificação da existência da mesma.

2.3 Trabalho relacionado

A pesquisa bibliográfica apresentada de seguida teve como objetivo melhor entender o conceito de pegada digital, o impacto da mesma sobre o utilizador e a sua relação com a privacidade.

Digital Footprinting: Uncovering Tourists with User-Generated Content

Girardin et al refere como a pegada digital passiva pôde ser utilizada para analisar a presença de turistas na cidade de Roma [14]. Esta informação fora obtida através de uma parceria com uma operadora de telecomunicações de modo a registar os eventos gerados pelo envio de SMS ou chamadas telefónicas. Este artigo levanta assim a atenção para um tipo de informação além do âmbito desta solução, os dados passivos. Estes são por norma guardados para a avaliação dos sistemas de informação.

Digital footprint in Web 3.0: Social Media Usage in Recruitment

O artigo por Nargesian et al [21] refere a entrevista a 10 gestores de recursos humanos sobre a sua perceção de pegada digital e a influência da mesma num processo de recrutamento. No artigo é possível concluir que os recrutadores utilizam a informação disponível acerca de um candidato no processo de decisão final. Os resultados das entrevistas revelam que os entrevistadores validam a presença em redes sociais dos entrevistados em busca de razões para não contratar os mesmos – isto poderá indicar que a presença em redes sociais é por si só um fator negativo no processo de recrutamento. Não obstante, o conceito de pegada digital contribui para o aumento da complexidade do processo de recrutamento.

Digital Footprints: Your Unique Identity

Blue et al propõem uma solução que procura utilizar a pegada digital de um individuo como forma de criar identidades *online* a serem utilizadas em sistemas de verificação humana [40]. Este trabalho distingue os conceitos de pegada e identidade digital classificando este último como uma complexa interação de diversos indivíduos e sistemas por forma a criar uma reputação *online*.

Assim sendo, a pegada digital atua como evidência de uma identidade digital. Esta solução estende a utilidade da gestão da pegada digital além do contexto de reputação.

Rethinking the *Web* as a Personal Archive

O artigo de Lindley et al incide sobre uma solução para a gestão da informação pessoal *online* onde propõem a divisão da mesma em cinco subtipos [7]. No processo de entrevista deste trabalho é referido o abandono de plataformas *online*, dado pelo coletivo abandono do círculo social do utilizador. Esta propensão para o abandono de plataformas reforça a necessidade de uma sensibilização sobre a pegada digital pois será fácil um utilizador esquecer-se dos registos que tem.

Numa discussão sobre que tipo de informação gostariam de manter *online* alguns dos entrevistados revelam que não têm interesse em guardar informação das redes sociais Twitter e Facebook devido à forma despreocupada como utilizam as mesmas. Apesar de o Twitter e Facebook serem atualmente das redes sociais do tipo blog mais utilizadas até à data esta observação indica que os utilizadores estarão interessados em reduzir a sua presença em plataformas semelhantes.

Unpacking “Privacy” for a Networked World

Leysia Palen et al [6] apresentam um modelo com o objetivo de relacionar a tecnologia com a gestão da privacidade de um indivíduo. Os autores baseiam-se na definição de privacidade do psicólogo Irwin Altman que a afirma como um conjunto seletivo de formas de acesso ao indivíduo dada por mecanismos comportamentais e processos dinâmicos e dialéticos. Isto implica a existência de tensões no ato de divulgação de informação pessoal sobre a forma como o indivíduo pretende apresentar-se ao mundo e como outros poderão utilizar a informação divulgada. Este conceito revela a intenção de uma pessoa em cuidar da sua imagem pública pelo que o mesmo pode ser feito através da gestão da pegada digital, contudo, o trabalho refere como a persistência da informação *online* dificulta este cuidado, em referência à persistência de publicações em fóruns antigos.

Podemos assim relacionar o conceito de privacidade com o da pegada digital no sentido de que esta complementa a privacidade permitindo ao utilizador ter algum controlo sobre a forma como quer ser percecionado.

Capítulo 3

Avaliação da necessidade de ferramentas da pegada digital

Foi realizado um questionário com o objetivo de fundamentar a utilidade de uma solução para a gestão da pegada digital e priorizar os tipos de plataformas *web* a serem pesquisadas, conforme descrito no processo do Capítulo 4. Inicialmente foi recolhida informação acerca da idade e tempo de utilização da *internet*, útil para correlacionar com os serviços mais importantes a pesquisar. De seguida, colocaram-se questões que permitiram determinar o rácio entre o número de contas conhecidas e de quantas o utilizador mantém registo por modo a entender a perceção do utilizador sobre o controlo da sua presença *online*. Por fim recolheu-se informação acerca do conhecimento sobre ferramentas que auxiliam a gestão da pegada digital, o nível de familiaridade com o tema e qual a preferência sobre os tipos de plataformas *web* a descobrir. Neste capítulo será apresentada a estrutura do questionário seguida de uma discussão sobre os resultados obtidos.

3.1 Questionário de avaliação da necessidade

O questionário referido obteve um total de 92 respostas e manteve-se ativo durante 30 dias após os quais se procedeu à análise dos resultados.

Faixa-etária

Questão original: How old are you?

Tabela 3.1 - Resultados sobre a faixa etária

Resposta possível	Resultado
< 18 anos	1% (1)
18 a 35 anos	75% (69)
36 a 55 anos	15% (14)
> 55 anos	8% (8)

Foi possível observar que grande parte do universo se encontra na faixa etária referente a adultos e jovens adultos. Contudo, deverá ter-se em conta que o questionário foi divulgado entre a comunidade de alunos da Faculdade de Ciências da Universidade de Lisboa e os círculos sociais do autor.

Utilização da *internet*

Questão original: How long have you been using the internet for?

Tabela 3.2 - Intervalo de utilização da internet, em anos

Resposta possível	Resultado
< 3 anos	0
3 a 10 anos	14% (13)
10 a 20 anos	69% (63)
20 a 40 anos	17% (16)

Como esperado através das idades aferidas na questão anterior, grande parte dos resultados recaem sobre o intervalo de 10 a 20 anos. Deverá ainda dar-se alguma atenção ao número de utilizadores presentes na *internet* há mais de 20 anos pois estes poderão dar um *feedback* baseado numa maior experiência nas questões que se seguem.

Nº de contas na *internet*

Questão original: How many accounts do you believe you have online?

Tabela 3.3 - Nº contas online, por utilizador

Resposta possível	Resultado
< 10 contas	16% (15)
10 a 30 contas	31% (29)
30 a 50 contas	25% (23)
> 50 contas	27% (25)

Não existe uma tendência para qualquer intervalo de contas de utilizador.

Nº de contas sob as quais o utilizador mantém registo

Questão original: How many online accounts do you keep track of? Either in a notepad or simply written down in a piece of paper.

Tabela 3.4 - Nº de contas sob registo

Resposta possível	Resultado
< 10 contas	47% (44)
10 a 30 contas	40% (37)
30 a 50 contas	5% (5)
> 50 contas	6% (6)

Para efeitos de registo considerou-se todo o tipo de formatos, físicos ou digitais. Os dados obtidos indicam que existem utilizadores que não possuem registo de todas as contas que percecionam ter *online*.

Conhecimento sobre as ferramentas de redução da pegada digital “HaveIBeenPwned” e “Unroll Me”

Questão original: Are the services “HaveIBeenPwned” or “Unroll.me” familiar to you?

Tabela 3.5 – Resposta sobre as ferramentas HaveIBeenPwned e Unroll Me

Resposta possível	Resultado
Nada familiar	50% (46)
Pouco familiar	9% (9)
Algo familiar	12% (11)
Muito familiar	9% (9)
Extremamente familiar	18% (17)

As ferramentas referidas na questão, tal como apresentadas no Capítulo 2, são utilizadas na redução da pegada digital. Foi possível observar que metade dos entrevistados não está de todo familiarizado com este tipo de ferramentas.

Conhecimento sobre o tema “Pegada Digital”

Questão original: Is the term “Digital footprint” familiar to you?

Tabela 3.6 - Familiaridade com o tema "Pegada Digital"

Resposta possível	Resultado
Nada familiar	7% (7)
Pouco familiar	15% (14)
Algo familiar	23% (22)
Muito familiar	30% (28)
Extremamente familiar	22% (21)

Os resultados obtidos nesta questão revelam que parte do público-alvo sabe ou já ouviu falar sobre o conceito de “pegada digital”, conforme descrito no Capítulo 2.

Necessidade de uma solução para a descoberta de contas *online* do utilizador e eventual remoção das mesmas

Questão original: How useful would you consider a service that discovers all accounts you may have online and help you delete them if needed?

Tabela 3.7 - Nível de necessidade de uma solução de redução da pegada digital

Resposta possível	Resultado
Nada importante	4% (4)
Pouco importante	0
Algo importante	19% (18)
Muito importante	34% (32)
Extremamente importante	41% (38)

Esta questão procurou avaliar a necessidade para uma solução como aquela apresentada neste trabalho.

Preferência sobre o tipo de contas *online* a remover

Questão original: What type of online accounts would you delete if you could? You may select multiple options.

Tabela 3.8 - Tipos de conta mais relevantes

Resposta possível	Resultado
Lojas	76% (70)
Fóruns <i>web</i>	71% (66)
Jogos	64% (59)
Redes Sociais	62% (57)
Outras	8% (8)

Esta questão permitia a seleção de múltiplas respostas e ainda opção de resposta aberta em “Outras” da qual se obtiveram as seguintes:

- *Websites* com pouca segurança que divulgam a informação do utilizador;
- Serviços *online* esquecidos pelo utilizador.

Discussão

Os resultados obtidos permitem-nos apurar que mais de dois terços da população-alvo deste questionário, maioritariamente situada entre os 18 e 35 anos, utiliza a *internet* há mais de uma década. A relação entre a idade dos utilizadores e a sua resposta à preferência entre o tipo de contas a remover revela que os 23% que pertencem à faixa etária superior aos 35 anos têm como preferência eliminar contas do tipo fóruns *web* e lojas. Por outro lado, os utilizadores mais jovens têm preferência sobre redes sociais e também lojas.

Através do cálculo da média das respostas sobre o número de contas que o utilizador percebe ter e a média sobre o número de contas sobre as quais o mesmo tem registo foi possível entender que o utilizador mantém em média registo de apenas 57% das suas contas na *internet*.

Cada utilizador possui, portanto, uma média de registo de 18 contas face o total de 32 que percebe a ter. Este resultado visa complementar a necessidade da criação de uma solução para a gestão da pegada digital.

Note-se de seguida o conhecimento geral dos utilizadores sobre o tema da pegada digital e soluções para a gestão da mesma (Tabela 3.5 e Tabela 3.6, respetivamente). Este indicador poderá relacionar-se com a aceitação geral da necessidade de uma solução para a gestão da pegada digital (Tabela 3.7) pois foi possível analisar que os 4 questionados que afirmam não ser importante uma solução para a gestão da pegada digital encontram-se entre os menos familiarizados com o conceito.

A forma de pesquisar o registo de contas nas diferentes plataformas *web* envolve a exploração manual e o desenvolvimento de uma base de dados de conhecimento que irá conter a biblioteca dos *websites* a serem verificados. Deste modo, o universo de *websites* sob análise irá inicialmente reger-se de acordo com as respostas obtidas na questão sobre a preferência de contas *online* a remover por forma a facilitar o trabalho manual de preenchimento da base de dados. Num trabalho futuro será interessante voltar à análise de correlação com idades por forma a reestruturar a prioridade aqui definida.

Capítulo 4

Desenho e implementação da solução

Ao longo deste capítulo é apresentada a solução proposta. No primeiro subcapítulo, dedicado ao desenho da solução, é feito um levantamento de requisitos e definido o âmbito sob o qual a mesma será desenvolvida.

O subcapítulo dedicado à implementação refere a forma como os requisitos previamente identificados foram desenvolvidos. Neste será descrita a arquitetura da solução, os processos criados e a estrutura de ambas as componentes de *frontend* e *backend*.

4.1 Desenho da solução

4.1.1 Análise de requisitos

De forma a alcançar os objetivos definidos foi feito um levantamento de requisitos funcionais e não-funcionais. Entre os não-funcionais encontram-se:

1. Política de privacidade – deverá ser claro para ambas as partes que a informação recolhida e processada pela solução não será partilhada com nenhuma outra entidade exceto quando permitida pelo utilizador-final para efeitos de comunicação e apelo à remoção da conta de utilizador. Esta política encontra-se sob o Regulamento Geral de Proteção de Dados.

Na vertente funcional estão incluídos:

2. Recolha de informação – a recolha será feita através de um formulário numa aplicação *web* e deve incluir obrigatoriamente o *e-mail* do utilizador-final. Após processada e obtidos os resultados, toda a informação deverá ser eliminada tanto da memória como do armazenamento local do(s) servidor(es) que a contém, com um período de expiração de quinze minutos, intervalo de tempo idealizado como suficiente para uma primeira análise de resultados. Este período poderá ser estendido através do prolongamento de uma *token* de sessão JWT (*JSON Web Token*).
3. Infraestrutura – o sistema que alojará a solução deverá suportar as tecnologias Python, Apache e um sistema de base de dados SQL. A solução poderá ser instalada em ambientes baseados em Windows ou Linux.
4. Base de dados de conhecimento – a solução deverá conter uma base de dados onde constem os processos de remoção da conta de utilizador para cada uma das plataformas-alvo identificadas. Adicionalmente, deverá ter a capacidade de identificar o país onde são processados da aplicação *web* sob análise e apresentar o *e-mail* formal correspondente. Estes processos deverão ser revistos periodicamente assegurando que continuam a servir o seu propósito.
5. Processamento de dados – o processamento da informação deverá ser conseguido através da linguagem de programação Python. A escolha desta deve-se ao extenso

repositório de bibliotecas auxiliares que irão compor a solução final. De forma a garantir uma fácil manutenção da solução, deverão ser utilizadas bibliotecas atualizadas.

4.1.2 Aplicação das leis de privacidade

Como referido anteriormente, no processo de remoção de conta deverá ser disponibilizada a opção de envio de um pedido formal via *e-mail*. Este deverá referenciar as leis de privacidade aplicáveis ao país de onde os dados da plataforma são guardados e ou processados. Como tal, na comunicação enviada serão incluídas referências a, pelo menos, as seguintes leis e regulamentos:

- CCPA – *California Consumer Privacy Act*, aplicável a empresas que processem dados nos **EUA**;
- CalOPPA – *The California Online Privacy Protection Act*, aplicável a utilizadores no estado da **Califórnia**;
- PIPEDA – *The Personal Information Protection and Electronic Documents Act*, aplicável aos dados processados no **Canada**;
- GDPR – *General Data Protection Regulation*, aplicável aos dados processados dentro da **UE**;
- ePrivacy – *ePrivacy Directive & Regulation*, aplicável à utilização de *cookies* dentro da **UE**;
- *Cyber Security Law*, lei utilizada na **China** sobre a proteção de dados;
- *The Privacy Act 1998*, aplicável aos dados processados na **Austrália**.

A decisão de abordar as leis previamente mencionadas baseia-se na vontade de abranger um número limitado de países que possuam um nível elevado de regulação e aplicação das mesmas conforme avaliado pela DLA Piper [41]. Isto implica que o envio do *e-mail* formal encontra-se limitado, numa fase inicial da solução, às organizações cujos dados sejam processados em: EUA, Canada, EU, China e Austrália. Contudo, mantem-se como objetivo futuro expandir esta análise a países cuja avaliação se encontre nos níveis “robusto”, “moderado” ou “limitado” [41].

4.1.3 Âmbito de aplicação da solução

O universo de aplicações e plataformas *web* disponíveis *online* é extenso. Não sendo viável cobrir todas as plataformas-alvo existentes, o alcance da solução deverá ser segmentado por tipo de plataforma ou serviço *web*. De acordo com o questionário de avaliação de necessidade apresentado no capítulo 3, a prioridade sobre os tipos de contas *online* a analisar (Tabela 3.8) será:

1. Lojas *online*;
2. Fóruns da *internet*;
3. Contas de jogos;
4. Redes sociais;
5. Restantes serviços.

Serão incluídos num âmbito inicial os *websites* presentes nas bases de dados das ferramentas existentes e aqueles mais populares dentro das categorias desta listagem, obedecendo aos métodos de validação descritos no subcapítulo 4.2.

4.1.4 Aplicação da solução

A solução será aplicada em auditorias de segurança com o objetivo de auxiliar na fase de identificação de endereços de e-mail [42] e nos processos de sensibilização para a temática do *phishing*. Esta solução não deverá ser utilizada publicamente até nota em contrário pela Willway. Sempre que possível, a solução deverá apresentar informação sobre a pegada digital com o objetivo de sensibilizar os utilizadores sobre o tema.

4.1.5 Esquematização de processos

Da análise de requisitos resultou a criação dos processos listados abaixo e apresentados posteriormente sob diagramas de atividade.

- Recolha e processamento da informação;
- Gestão da sessão do utilizador;
- Pesquisa de contas de utilizador;
- Classificação da pegada digital;
- Gestão da informação;
- Gestão da base de dados de conhecimento.

Recolha e processamento da informação

No diagrama referente à recolha e processamento de informação (Figura 4.1) estão presentes três atores: o utilizador, responsável por iniciar o processo através da introdução de dados, o *frontend*, responsável por apresentar os resultados da pesquisa de forma gráfica e o *backend*, responsável pela componente técnica de gestão da sessão do utilizador, pesquisa sobre a pegada digital e classificação da mesma.

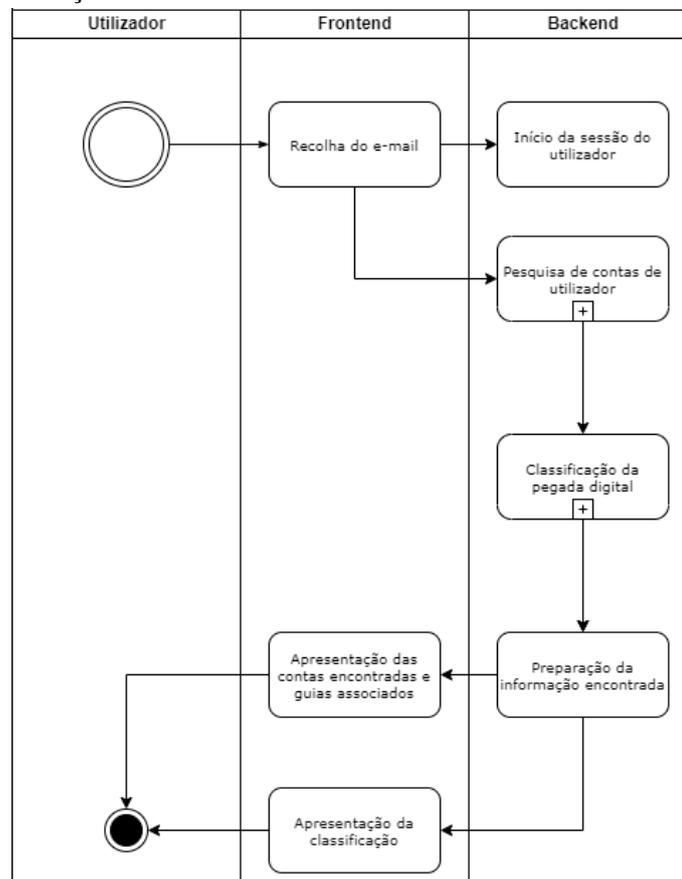


Figura 4.1 - Diagrama sobre a recolha e processamento de informação

O processo inicia-se com a recolha do *e-mail* através de um formulário *web*. Esta informação é então enviada para a base de dados onde será guardada por um período mínimo de quinze minutos. Com este registo inicia-se também o processo de gestão de sessão do utilizador. Paralelamente é iniciado o processo de pesquisa de contas de utilizador que irá validar onde existe um registo entre os diferentes *websites* na base de dados de conhecimento. Quando concluída a pesquisa dá-se a classificação da pegada digital que será apresentada ao utilizador pelo *frontend*. Adicionalmente o *backend* irá processar os resultados obtidos da pesquisa para devolver ao utilizador os guias gráficos e *e-mail* modelo necessário para a eventual remoção das contas encontradas.

Gestão da sessão do utilizador

O processo para a gestão da sessão do utilizador (Figura 4.2) segue a implementação de uma *silent session refresh* [43] através do uso de tokens JWT.

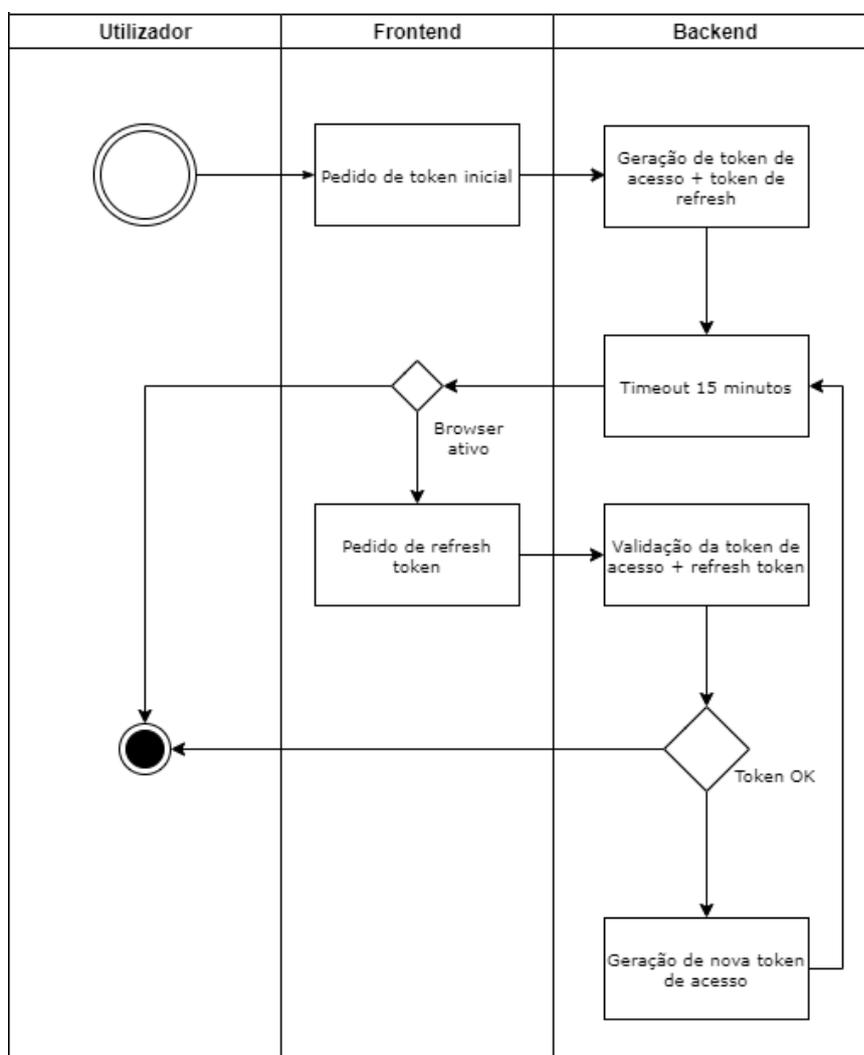


Figura 4.2 - Diagrama sobre a gestão da sessão do utilizador

Após a componente *frontend* efetuar o pedido de pesquisa serão geradas duas *tokens* de sessão: uma utilizada no acesso aos resultados da pesquisa e outra utilizada para o pedido de prolongamento da sessão. Após criadas e registadas na base de dados, é atribuída à *token* de acesso uma validade de quinze minutos. Após catorze minutos, se o utilizador mantiver o browser aberto,

o *frontend* iniciará um novo pedido de *token* de acesso, utilizando a *refresh token* para obter uma nova *token* de acesso durante outros catorze minutos. Contudo, se o algoritmo de validação falhar, a sessão do utilizador expirará e será encerrada com um código 401, devolvendo um redireccionamento para a página inicial.

A *refresh token* possui também uma validade de 30 dias, o que significa que o utilizador poderá, no máximo, manter o seu navegador aberto e continuar a prolongar a sua sessão por 30 dias. O processo de invalidação de *tokens* será discutido na secção de “gestão da informação”.

Pesquisa de contas de utilizador

A pesquisa de contas do utilizador (Figura 4.3) incide sobre uma base de dados de conhecimento que agrega um conjunto de *websites* utilizados pelas ferramentas existentes e aqueles mais utilizados dentro das categorias identificadas no questionário de avaliação. O processo de pesquisa assenta sobre a *framework web* Quart que permite que as pesquisas possam ser realizadas de forma assíncrona.

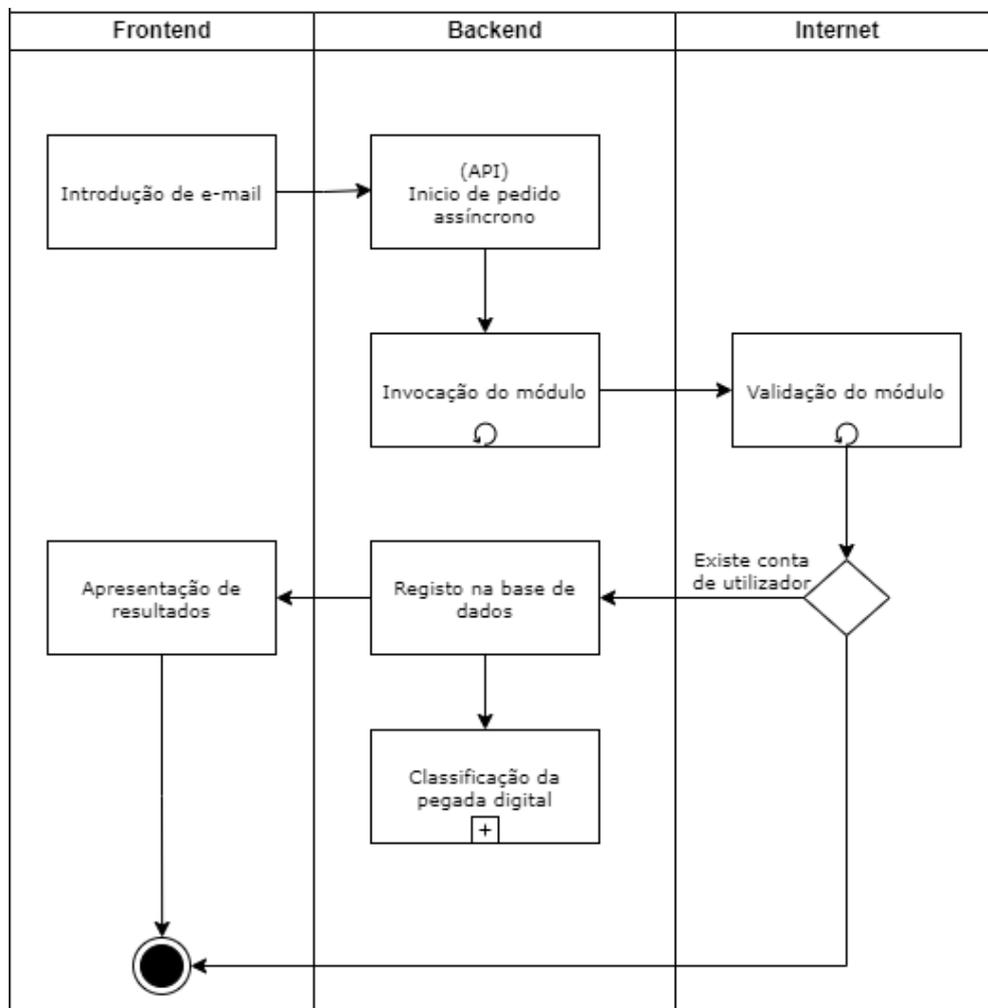


Figura 4.3 - Diagrama sobre a pesquisa de contas do utilizador

Um módulo, como referido na figura, é a representação técnica de um *website*, dentro de cada pesquisa, o *backend* irá validar cada módulo através da resposta do servidor *web* de destino sobre as tentativas de registo, *login* ou recuperação de *password* utilizando o *e-mail* do utilizador.

A cada pesquisa será associado um código identificador único e temporário posteriormente registado na base de dados em conjunto com cada um dos módulos onde a validação tenha sido positiva. Os resultados positivos serão utilizados no processo da classificação da pegada digital e devolvidos ao utilizador pelo *frontend*.

Classificação da pegada digital

O cálculo da classificação da pegada digital é feito no *backend* (Figura 4.4) através da soma do valor de cada módulo validado de forma positiva. A cada um destes é atribuído um campo de “tipo de *website*” como por exemplo “fórum”, “loja *online*” ou “rede social” e a cada tipo de *website* é atribuído um peso.

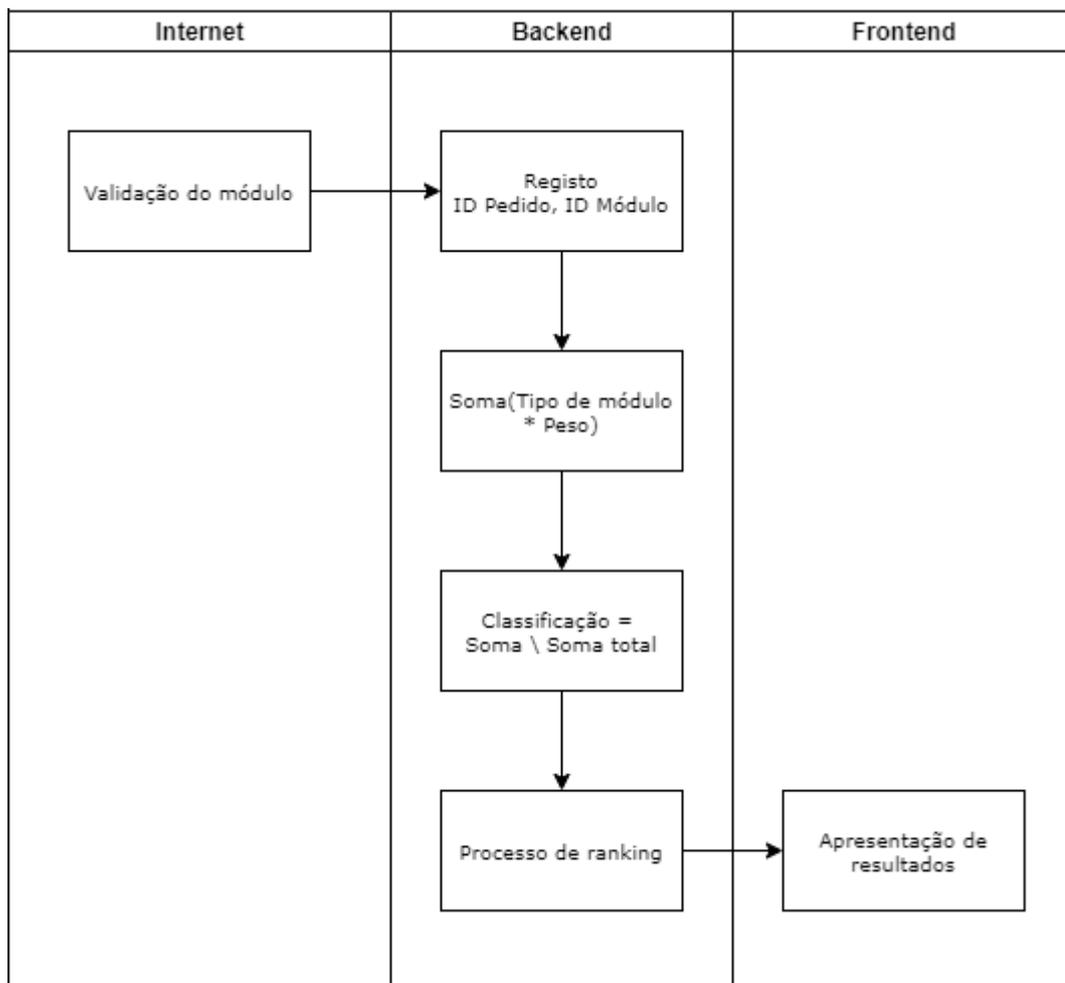


Figura 4.4 - Diagrama sobre a classificação da pegada digital

O peso atribuído a cada módulo baseia-se na prioridade definida na fase da avaliação de necessidade e foi distribuído de forma gradual consoante a mesma:

Tabela 4.1 - Pesos atribuídos por tipo de *website*

Tipo de <i>website</i>	Peso
Lojas <i>online</i>	1.5
Fóruns da <i>internet</i>	1.4
Contas de jogos	1.3
Redes sociais	1.2
Restantes serviços	1.1

A utilização de pesos (Tabela 4.1) tem como objetivo permitir que a solução acompanhe tendências sobre o tipo de *websites* mais utilizados ou outras alterações ao âmbito de aplicação.

A classificação final é dada pela soma dos valores associados aos *websites* onde tenha sido possível encontrar registo, dividida pelo valor máximo possível:

$$\text{Classificação} = \frac{\text{soma}(\text{website}())}{\text{máxima}}$$

Onde

$$\text{website}() = 1 * (\text{peso do tipo de website encontrado})$$

E

$$\text{máxima} = \text{classificação máxima resultante da soma de todos os websites}$$

Assim sendo, este cálculo resultará num valor percentual que será por fim utilizado para categorizar a pegada digital em 5 níveis:

Tabela 4.2 - Níveis de classificação da pegada digital

Classificação	Intervalo %
Muito elevada	80 - 100
Elevada	60 - 80
Moderada	40 - 60
Baixa	20 - 40
Muito baixa	0 - 20

Conforme se verificar o decréscimo do número de contas de utilizador, menor será a classificação da pegada digital do mesmo (Tabela 4.2).

Gestão da informação

A Figura 4.5 refere-se ao diagrama de atividade sobre o processo de gestão da informação recolhida na pesquisa da pegada digital

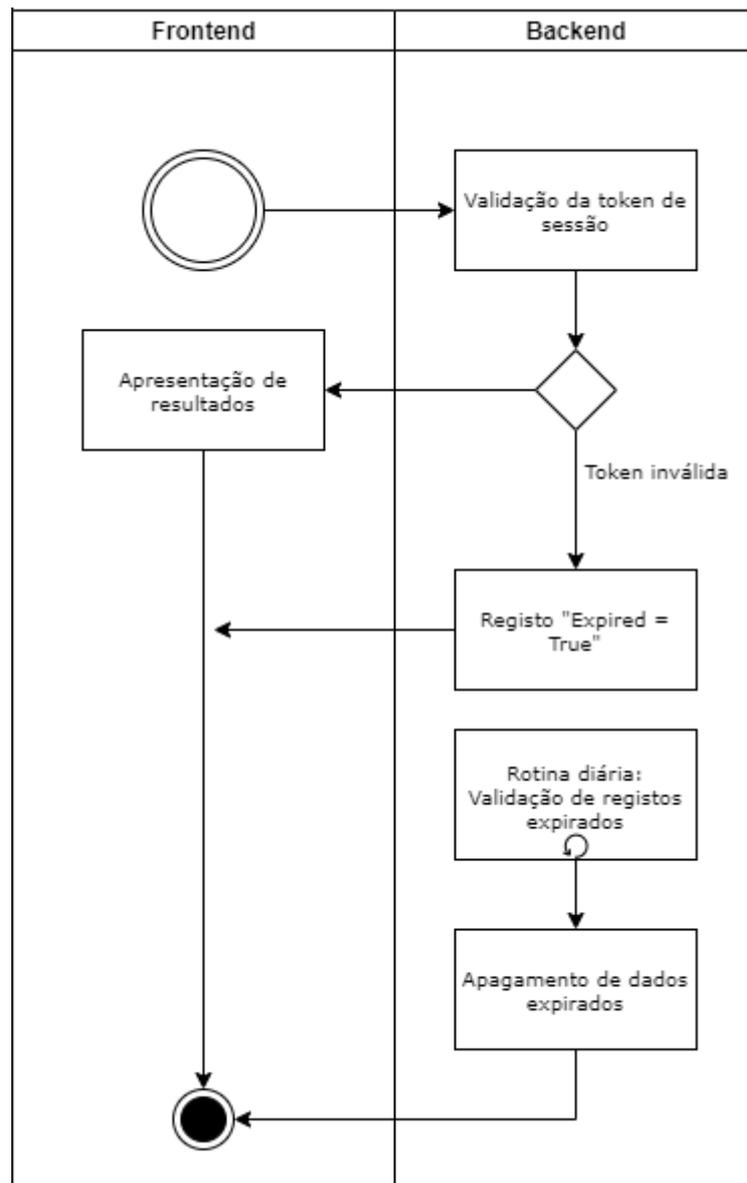


Figura 4.5 - Diagrama sobre a gestão da informação

. Quando uma *token* de sessão é expirada, ou seja, a data atual excede a data-limite registada na base de dados, é associado um registo de “expirado”. Este registo será posteriormente utilizado numa rotina diária da base de dados que tem como objetivo apagar todos os registos expirados eliminando assim as pesquisas dos utilizadores.

Gestão da base de dados de conhecimento

A frequência de atualização de *website* pode variar, seja por gosto pessoal, alterações ao *branding*, entre outras razões [44]. Dada a incerteza desta frequência foi estabelecido um intervalo de revisão de 30 dias.

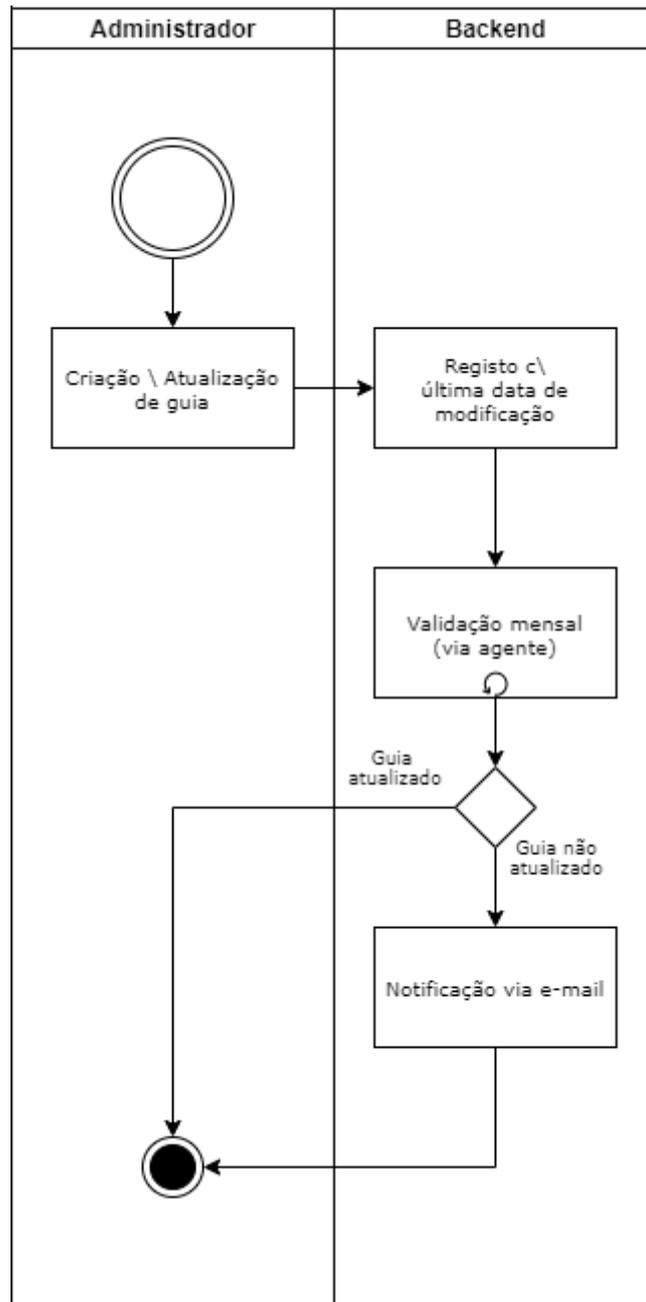


Figura 4.6 - Diagrama sobre a gestão da base de dados de conhecimento

O processo referente à gestão da base de dados de conhecimento (Figura 4.6) tem como objetivo garantir que os guias de remoção de conta se mantêm atualizados. Para tal, é feita uma validação mensal da última alteração de cada guia de interface com o objetivo de informar o administrador sobre os guias que necessitem de revisão.

4.2 Implementação da solução

De seguida será apresentada a estrutura técnica da solução e a implementação dos processos definidos na fase de desenho, respondendo aos requisitos previamente identificados. A solução terá o nome “Untrace”.

4.2.1 Política de privacidade

A política de privacidade garante a utilização dos dados necessários para o funcionamento e foi redigida de acordo com a atual política da Willway. Esta política encontra-se acessível a qualquer altura na plataforma *web* da Untrace através do painel principal de navegação e Anexo I – Política de privacidade.

O principal objetivo é garantir que os utilizadores são informados acerca de como a privacidade dos seus dados é garantida.

4.2.2 Infraestrutura de suporte

A implementação da infraestrutura da solução (Figura 4.7) envolveu a introdução de medidas de segurança dado que se encontra alojada na infraestrutura tecnológica da Willway. De modo a garantir que a solução é utilizada externamente pela Willway de forma segura foi imposto o requisito de um acesso VPN (*Virtual Private Network*). O acesso VPN tem como objetivo proteger a ligação entre um colaborador da Willway fora das instalações e a solução desenvolvida. A VPN requer credenciais de autenticação que, se corretamente validadas, dão acesso ao colaborador à rede onde se encontra a solução.

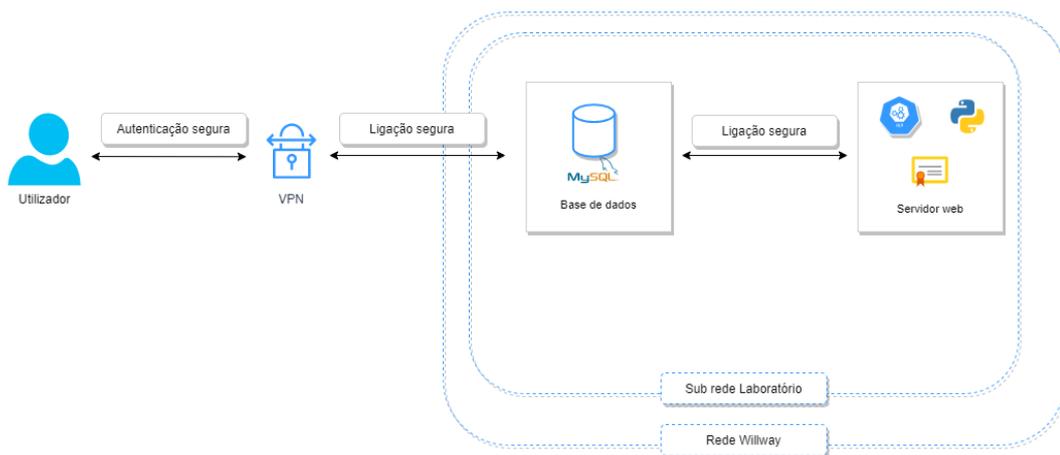


Figura 4.7 - Diagrama da infraestrutura

A base de dados encontra-se isolada num servidor da Willway com o intuito de mitigar um cenário de comprometimento do servidor da componente *web* em que um utilizador mal-intencionado procure aceder à base de dados associada. Como medida de segurança adicional, as credenciais do ficheiro de configuração para a ligação entre aplicação *web* e base de dados encontram-se cifradas.

Todas as passwords utilizadas no acesso às diferentes componentes *web* e base de dados foram geradas por um software de gestão de passwords, sendo as mesmas consideradas fortes [45].

Por último, a aplicação *web* encontra-se implementada sobre um protocolo de comunicação seguro (HTTPS), utilizando um certificado adquirido e autorizado por uma entidade certificadora reconhecida [46].

Tecnologia utilizada

De seguida será discutida e justificada a tecnologia utilizada (Figura 4.8) em cada componente da solução.

Frontend

A componente *frontend* tem como objetivo interligar o utilizador e os processos de gestão da pegada digital da solução através de uma interface gráfica. Existem diversas *frameworks* que cumprem o propósito sendo as mais populares [47] React.js, jQuery, Express.js, Angular e Vue.js.

Dado que a complexidade da solução se encontra no *backend*, para a construção *frontend* procurou determinar-se qual a mais computacionalmente leve, despida de bibliotecas [48], adaptável [49] e de fácil aprendizagem considerando o intervalo de tempo alocado ao desenvolvimento. Para medir estes indicadores foi realizado um exercício *web*, utilizando cada uma das *frameworks* referidas, que consistiu em construir uma versão base da página principal da Untrace e integrá-la com uma base de dados. Da avaliação deste exercício resultou a seguinte análise:

Tabela 4.3 - Comparação de *frameworks web*

Tecnologia <i>frontend</i>	Aspetos positivos	Aspetos negativos
React.js	Rapidez, menor complexidade através da abstração do DOM (virtual DOM). Escalável, através de código modular. Fácil integração com APIs <i>backend</i> .	Bibliotecas auxiliares. CLI. JSX. Atualizações frequentes.
jQuery	Leve. Documentação extensiva. JavaScript nativo. Compatibilidade com navegadores <i>web</i> .	Os navegadores mais recentes já implementam partes da API.
Express.js	JavaScript nativo. Assíncrono.	Node.js.
Angular	Escalável através de código modular. Fácil integração com APIs <i>backend</i> .	Bibliotecas auxiliares. CLI. Typescript. Atualizações frequentes. Arquitetura MVC.
Vue.js	Leve. Sintaxe simples. Rapidez, menor complexidade através da abstração do DOM (virtual DOM).	Arquitetura MVVW. Barreira linguística.

React.js [50] é uma biblioteca de JavaScript que implementa uma virtualização do DOM (*Document Object Model*), interface utilizada na modelação de páginas *web*, de modo a reduzir a complexidade computacional de atualizar a mesma quando uma alteração é necessária, por exemplo em eventos de clique. A vertente modular de React permite, através de componentes, reutilizar código *web*, o que facilita a escalabilidade da solução. Adicionalmente, verificou-se ser de fácil integração com APIs e por sua vez a base de dados do exercício. Entre os aspetos

negativos desta *framework* encontram-se a necessidade de instalação de bibliotecas auxiliares, tornando a solução computacionalmente mais pesada, a necessidade de aprendizagem sobre a utilização de uma CLI (*Command Line Interface*) e uma linguagem de programação baseada em JavaScript, JSX (*JavaScript Syntax Extension*). O facto de ser uma biblioteca em constante atualização dificulta o processo de atualização de código e requer um constante acompanhamento da evolução da sintaxe do mesmo.

JQuery [51] mostrou ser a tecnologia *frontend* mais adequada para o projeto pois contém uma documentação extensiva [52], ocupa um mínimo total de 30KB [53], utiliza JavaScript como linguagem de programação base e oferece um intervalo considerável de compatibilidade entre navegadores *web* [54]. Contudo, as funcionalidades mais relevantes de JQuery já são atualmente possíveis com JavaScript puro [55].

Express [56] é uma *framework web* utilizada em conjunto com a *framework backend* Node.js que suporta nativamente a gestão de pedidos *web* assíncronos, úteis para a execução da pesquisa da pegada digital. Apesar de ser também baseada em código JavaScript puro esta *framework* está dependente da implementação do *backend* Node.js [57].

Angular [58] afigura-se como uma *framework* de características semelhantes a React quanto à sua escalabilidade e facilidade de integração com APIs, necessidade de bibliotecas que trazem algum peso à instalação, aprendizagem de uma CLI e uma linguagem de programação nova sob constantes atualizações que requerem um acompanhamento próximo da mesma. A sua complexidade é ainda aumentada pela utilização de uma arquitetura MVC (*Model-View-Container*) que tem como objetivo separar as componentes de interface, dados e lógica de negócio de uma aplicação *web*.

Vue.js [59] é uma *framework web* que à semelhança de React abstrai o DOM via o Virtual DOM e possui um tamanho altamente reduzido semelhante a JQuery. Esta utiliza a arquitetura MVVW (*Model-View-ViewModel*) que apesar de ser mais simples por não incorporar a lógica de negócio adiciona complexidade desnecessária a uma solução que tem como objetivo ser simples. Outro aspeto negativo é o da barreira linguística que se deve ao facto de Vue.js ser uma das *frameworks web* mais populares na China [60], resultando num esforço de tradução e compreensão elevados.

Backend

O objetivo desta componente é gerir a informação e executar os processos para a gestão da pegada digital. Na escolha da tecnologia *backend* procurou-se utilizar uma linguagem de *scripting* considerando que o objetivo da solução é automatizar um conjunto de tarefas já existentes, em contrário ao uso de uma linguagem de programação cujo objetivo seria criar essas mesmas tarefas. Por conseguinte, pretendeu-se utilizar uma tecnologia que automatizasse pedidos *web* e acessos a uma base dados preferencialmente de forma assíncrona de modo a reduzir o tempo de espera sobre a pesquisa da pegada digital.

Considerando que a Untrace utilizará uma adaptação de Holehe para a pesquisa de contas de utilizador, foi escolhida a linguagem de desenvolvimento Python para a criação de uma API (*Application Programming Interface*) que irá conectar os pedidos do utilizador aos seus respetivos processos. A criação de uma API implica por sua vez a escolha de uma *framework*, utilizada na criação do esqueleto da mesma.

A escolha de Quart [61] e Flask [62] para a implementação do processo de pesquisa da pegada digital (Figura 4.8) baseou-se na necessidade de paralelismo. Esta funcionalidade permite que a solução efetue pesquisas sobre a pegada digital em simultâneo providenciando ao utilizador resultados de forma mais rápida quando comparado com a alternativa que é a pesquisa em modo

sequencial. A escolha destas tecnologias baseou-se ainda na sua linguagem de desenvolvimento, Python [63] [64], a qual era a mais familiar aquando da escolha de uma linguagem adequada ao projeto.

A escolha sobre a utilização de MySQL como tecnologia de base de dados incide sobre o facto de a solução ter um processo de funcionamento estável. Não é esperado que o modo de pesquisa de contas de utilizador venha a alterar sendo que futuros desenvolvimentos focar-se-ão na sensibilização da pegada digital e gestão e criação dos atuais módulos de pesquisa. Deste modo é eliminada a necessidade de um sistema NoSQL [65]. Adicionalmente, a escolha entre MySQL e PostgreSQL foi meramente baseada na maior familiaridade com a primeira tecnologia.

Na Figura 4.8 encontra-se representada a arquitetura tecnológica discutida.

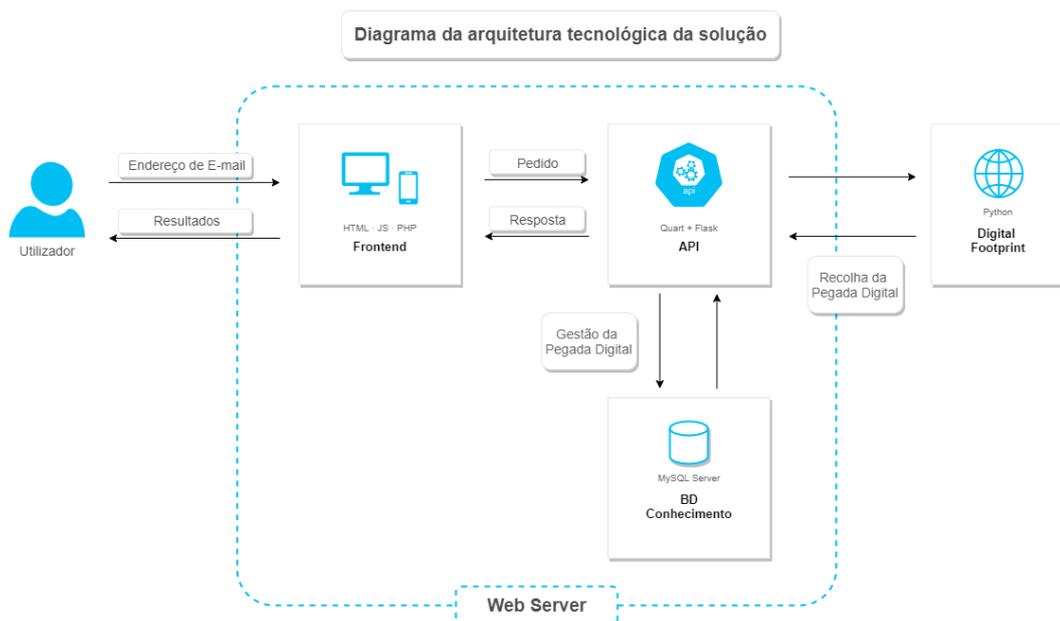


Figura 4.8 - Arquitetura tecnológica

4.2.3 Base de dados de conhecimento

Objetivo

Uma base de dados é uma estrutura de dados que contém informação essencial para o funcionamento da solução. Neste projeto esta toma o nome de “base de dados de conhecimento” devido à importância e valor da informação contida. Até à data a base de dados alberga:

- Listagem dos *websites*-alvo;
- Guias para a informação acerca do modo de remoção da conta via interface *web*;
- Modelo de *e-mail* para cancelamento de contas;
- Informação acerca das *tokens* de sessão do utilizador;
- Classificação dos diferentes tipos de *website*.

. O objetivo da base dados é fornecer informação atualizada sobre os guias de remoção de conta de utilizador.

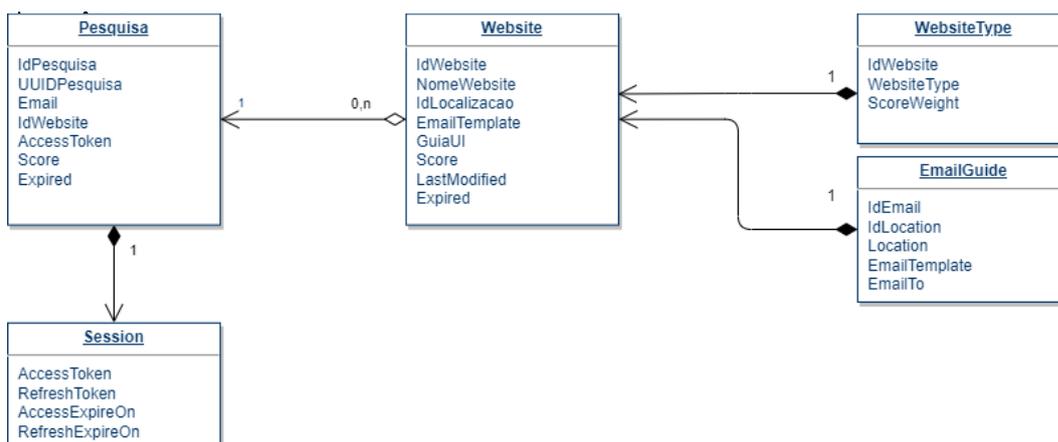


Figura 4.9 - Arquitetura da base de dados

O diagrama acima (Figura 4.9) representa a estrutura atual da base de dados. Os resultados testados como positivos serão guardados na tabela “Pesquisa” que por sua vez irá recorrer às tabelas “Session” para registo da sessão do utilizador e “Website” para apresentar a informação referente a cada tipo de *website*. A tabela “WebsiteType” é utilizada para o cálculo das classificações de cada tipo de *website* e poderá ser alterada conforme o peso dos diferentes tipos de conta seja alterado. Por fim, a tabela “EmailGuide” contém um *template* de e-mail a enviar consoante o país (localização) do processamento de dados tomando partido quando possível das leis de privacidade identificadas no subcapítulo de desenho da solução.

Classificação da pegada digital

A base de dados é responsável por manter um registo do nível de pureza da pegada digital do utilizador. Este nível de pureza é calculado numa escala de 0 a 100 baseada no rácio entre contas encontradas e aquelas conhecidas pelo requisitante. O cálculo é através de uma *query* SQL (Figura 4.10).

```
Score = 1 *
(
  SELECT scoreweight
  FROM   websitetype,
        website
  WHERE websitetype.idwebsite = website.idwebsite)
```

Figura 4.10 - Excerto da Query SQL

Por fim, para efeitos de apresentação, será feita a soma de todos os cálculos pertencentes à pesquisa efetuada. Da soma retiramos o valor percentual face o cálculo total possível. O *frontend* tratará de apresentar o texto classificativo correspondente (Tabela 4.2).

Gestão da informação

O processo de gestão da informação é responsável por remover os registos de pesquisa expirados. Como tal, foi criada o seguinte evento calendarizado em MySQL (Figura 4.11).

```
CREATE event expirecheck ON schedule every 24 hours starts '
2021-09-
09 00:01:00' comment 'Removes expired records older than 15 minu
tes' do
DELETE
FROM pesquisa
WHERE Expired = 1
OR
(
    SELECT session.AccessExpireOn
    FROM session,
    pesquisa
    WHERE session.AccessToken = pesquisa.AccessToken)
< Now()- interval 15 minute;
```

Figura 4.11 - Query SQL agendada

Este evento irá ser executado diariamente às 01:00 e visa dar resposta a dois cenários de expiração:

Cenário 1: Onde a pesquisa do utilizador tenha sido expirada pois o mesmo tentou obter uma prolongação da sessão passados pelo menos quinze minutos.

Cenário 2: Onde a pesquisa do utilizador se encontre expirada sem que o utilizador tenha tentado posteriormente obter uma prolongação da sessão.

Gestão da base de dados de conhecimento

A gestão dos guias referentes aos processos de remoção de contas é feita através de uma validação da tabela “Website” (Figura 4.9). Para tal, foi criada uma “Windows Scheduled Task” em Python que mensalmente acede à base de dados e valida a última data de modificação. Os registos que não tenham sido alterados nos últimos 30 dias farão parte de um email informativo a ser enviado para o administrador da solução.

```
with open('emailHTMLBuilt.html', 'r') as file :
filedata = file.read()
data = {
    "mail_settings": {

    },
    "personalizations": [
        {
            "to": [
                {
                    "email": "jose.martins@willway.pt"
                }
            ],
            "subject": "[Willway] Monthly Status Report"
        }
    ],
    "from": {
        "email": "no-reply@willway.pt",
        "name": 'Untrace Reporter'
    },
    "content": [
        {
            "type": "text/html",
            "value": filedata
        }
    ]
}
```

Figura 4.12 - Excerto do script de validação mensal

Neste caso, é possível observar que o *e-mail* final se encontra apenas endereçado ao endereço de *e-mail* jose.martins@willway.pt (Figura 4.12) com a possibilidade de incluir outros

utilizadores em CC. O envio do *e-mail* é feito através da integração da API do Sendgrid [66], um serviço utilizado para a entrega e monitorização de campanhas de *e-mail*.

Na construção do *e-mail* estarão incluídos apenas os guias que se encontrem desatualizados há pelo menos 30 dias. Na Figura 4.13 encontra-se um exemplo do *e-mail* enviado para o administrador.

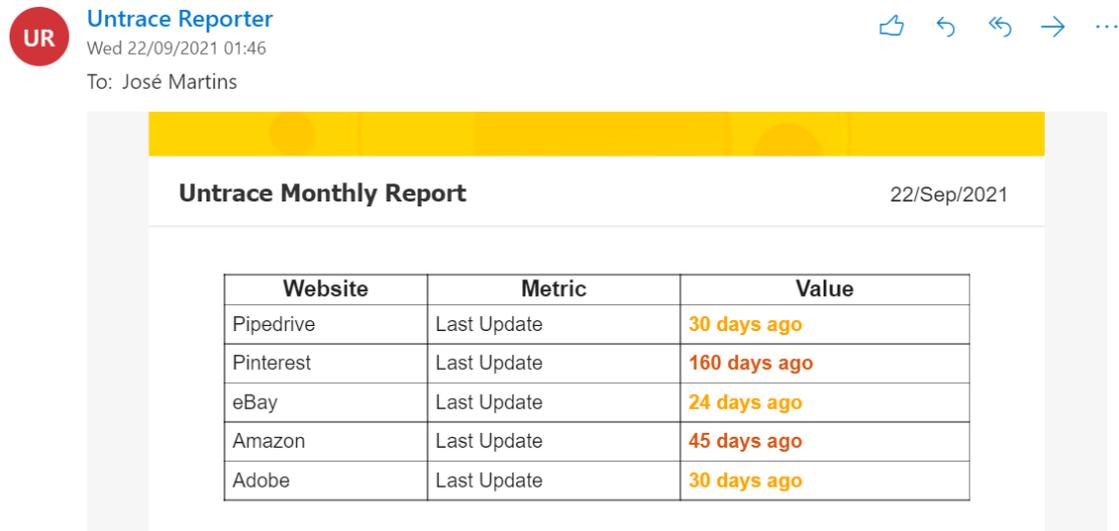


Figura 4.13 - *E-mail* exemplo de validação mensal

A estrutura do *e-mail*, construída em HTML, é gerada previamente de forma dinâmica através dos resultados de *queries SQL* executadas pelos módulos “MySQL Connector Python” e “PyMySQL”.

4.2.4 Componente *web*

Objetivo

O objetivo da componente *web* da solução é disponibilizar uma interface gráfica através da qual o utilizador poderá fazer a sua pesquisa e iniciar o processo de remoção das suas contas *online*.

Esta componente irá comunicar com o *backend* para o registo e consulta de informação da mesma.

Arquitetura



Figura 4.14 - Diagrama da componente *web*

Existem duas subcomponentes: o *frontend* e a API (Figura 4.14). O *frontend* é composto pelas seguintes páginas:

- *Homepage*;
Página que contém a funcionalidade de pesquisa;
- *Search*;
Página que irá conter os resultados da pesquisa com os guias associados;
- *Privacy Policy*;
Página com a referência à política de privacidade;
- *About*;
Página com informação sobre a pegada digital.

A API contém, por sua vez, *endpoints* utilizados nas seguintes funções:

- Pesquisa assíncrona de *e-mails*;
- Classificação da pegada digital;
- Gestão de sessão do utilizador;
- Envio de *e-mails*;
- Disponibilização de informação adicional.

Permissões

Ao concordar com a política de privacidade o utilizador dá o seu consentimento sobre a utilização do *e-mail* indicado para a pesquisa de conta de utilizador. Apesar desta pesquisa não incluir nenhum método que envolva o registo em itens do tipo “*newsletter*” ou divulgação de informação publicitária, não podemos garantir que todas as aplicações *web* onde é feita esta procura garantam o mesmo tratamento dos dados. Deste modo a permissão concedida poderá ter implicações fora do controlo da solução.

Recolha e processamento da informação

O processo de recolha de informação inicia-se na *homepage* (Figura 4.15). É através desta que o utilizador poderá utilizar o formulário de pesquisa inserindo o seu *e-mail*.

Através da *homepage* será ainda possível ler sobre o conceito da pegada digital e uma breve descrição do funcionamento da Untrace.

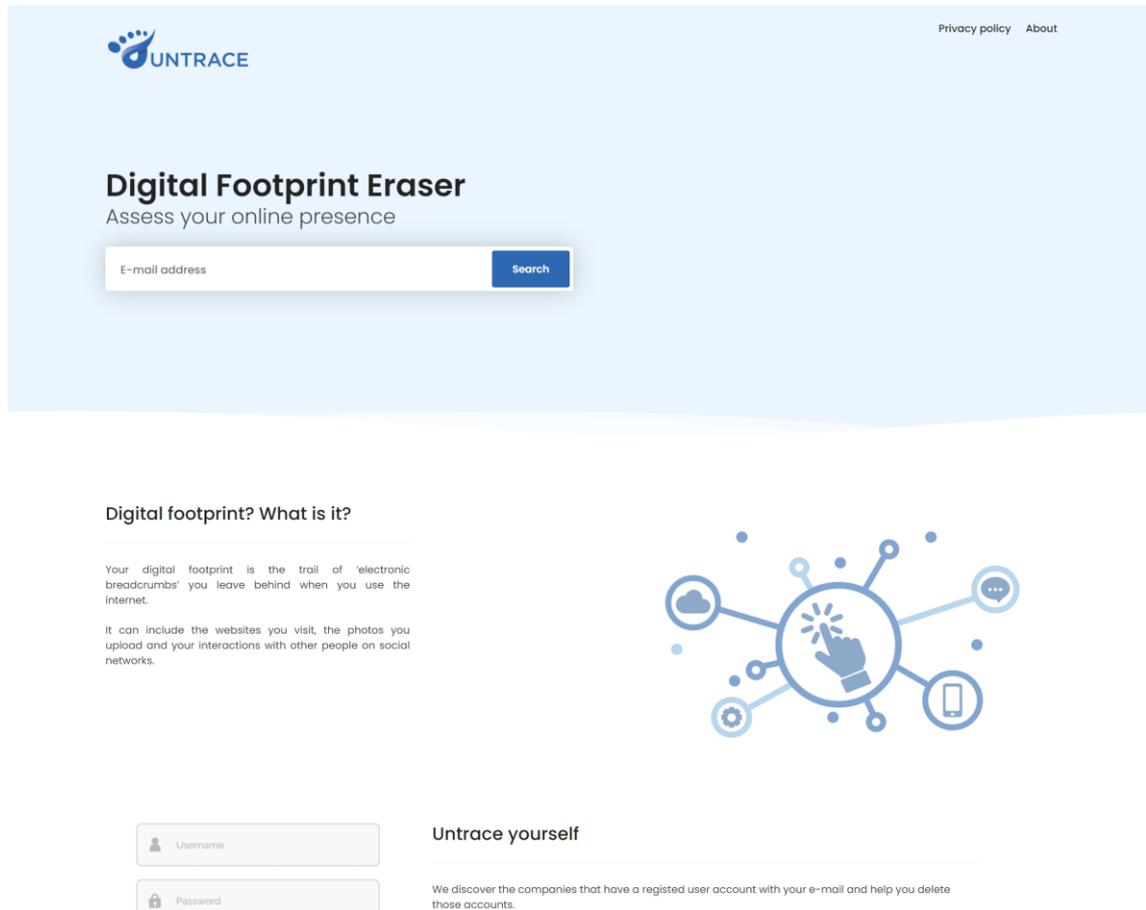


Figura 4.15 - Untrace – homepage

O utilizador poderá ainda, a qualquer momento, consultar a política de privacidade e mais informação sobre a solução acedendo às páginas “*Privacy Policy*” ou “*About*”.

Pesquisa de contas

O processo de pesquisa de contas associadas ao *e-mail* inicia após o formulário da *homepage* ser submetido. A Figura 4.16 ilustra este processo.

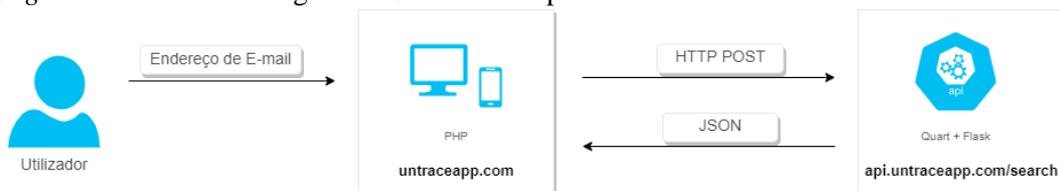


Figura 4.16 - Pedido de pesquisa de contas

Previamente ao início do pedido a estrutura do endereço de *e-mail* é validada por forma a impedir ataques de injeção de código. De seguida, é feito um pedido *POST* com a informação do formulário para o *endpoint* “*search*” da API da solução (Figura 4.17).

```

if (isset($_POST['email'])) {
    //POST data variable:
    $data = array('email' => $_POST["email"]);

    //API Call - returns a JSON decoded response
    $api_data = callAPI('POST', 'https://api.untraceapp.com/search', $data);

    for ($i = 0; $i < count($api_data); $i++) {

```

Figura 4.17 - Excerto de código referente ao pedido à API

Após realizado o pedido, a execução da pesquisa passa para o lado da API. Através da linguagem de programação Python e bibliotecas de assincronismo será feita a pesquisa em diversos *websites* utilizando o *e-mail* indicado. Até à data é possível efetuar a pesquisa nos *websites* referidos na Tabela 4.4.

Tabela 4.4 - Websites suportados

aboutme	ebay	tunefind	pinterest
adobe	ello	twitter	pipedrive
amazon	envato	venmo	plurk
amocrm	eventbrite	vivino	pornhub
anydo	evernote	voxxmedia	protonmail
archive	fanpop	vrbo	quora
armurerieauxerre	firefox	vsco	raidforums
atlassian	flickr	wattpad	rambler
axonaut	freelancer	wordpress	redtube
babeshows	freiberg	xing	replit
badeggsonline	garmin	xvideos	rocketreach
biosmods	github	yahoo	samsung
biotechnologyforums	nextpvr	google	seoclerks
bitmoji	nike	gravatar	sevincups
blablacar	nimble	hubspot	smule
blackworldforum	nocrm	imgur	snapchat
blip	nutshell	insightly	sporcle
blitzortung	odampublishing	instagram	spotify
bluegrassrivals	odnoklassniki	issuu	strava
bodybuilding	office365	koditv	taringa
buymeacoffee	onlinesequencer	komoot	teamleader
cambridgemt	parler	laposte	teamtreeshouse
caringbridge	crevado	lastfm	tellonym
chinaphonearena	deliveroo	lastpass	thecardboard
clashfarmer	demonforums	mail_ru	therianguide
codecademy	devrant	mybb	thevapingforum
codeigniter	diigo	myspace	treasureclassified
codepen	discord	nattyornot	s
coroflot	docker	naturabuy	tumblr
cpaelites	dominosfr	ndemiccreations	cracked_to
			cpahero

Caso se confirme a existência de uma conta de utilizador (Figura 4.18) em um ou mais dos *websites* referidos a informação é registada na base de dados para posterior consulta e apresentação ao utilizador que efetuou a pesquisa. A esta pesquisa será atribuído um código identificador interno que contém 48 caracteres alfanuméricos, de forma a prevenir ataques de *brute force*, e uma *token* de sessão.

```
req = await client.get(
    "https://www.pinterest.com/_ngjs/resource/EmailExistsResource/get/",
    params={
        "source_url": "/",
        "data": '{"options": {"email": "' + email + '"}, "context": {}}'})
```

Figura 4.18 - Excerto de código de validação de registo no website Pinterest

O modo de pesquisa de uma conta pode variar: certos *websites* podem já possuir alguma API disponível para esse propósito enquanto outros requerem que se envie um pedido de registo de conta para verificar se o *e-mail* existe nos seus sistemas (Figura 4.19).

```
headers["x-csrf-token"] = token
check = await client.post(
    "https://www.instagram.com/accounts/web_create_ajax/attempt/",
    data=data,
    headers=headers)
check = check.json()
if check["status"] != "fail":
    if 'email' in check["errors"].keys():
        if check["errors"]["email"][0]["code"] == "email_is_taken":
```

Figura 4.19 - Excerto de código de validação de registo no website Instagram

Remover uma conta

Após devolvida a pesquisa, é apresentado ao utilizador uma página *web* com os resultados da mesma. Esta página *web* (Figura 4.20) contém todos os *websites* onde foi identificado o *e-mail* e para cada qual duas ações disponíveis: remover a conta via interface *web* ou através do envio de um *e-mail*.

The screenshot shows the UNTRACE website interface. At the top, there is a navigation bar with the UNTRACE logo and links for 'Privacy policy' and 'About'. Below this, a large blue arc represents the digital footprint score, which is 14%, labeled as 'Very Low'. A text box explains that the score is based on the number of accounts found compared to the grand total and that it can be reduced by deleting found accounts. A list of websites where the user's email was found is displayed, including Pinterest, Amazon, Google, and Office365. Each website entry includes the logo, the name, and two buttons: 'Web GUI' and 'E-mail'.

Figura 4.20 - Página web de resultados, o e-mail encontra-se omitido

O utilizador poderá consultar a sua classificação da pegada digital através da página de resultados (Figura 4.21). De um modo geral, um menor número de contas de utilizador resultará numa classificação positiva da pegada digital. Este nível poderá ainda ser alterado conforme sejam encontradas mais ou menos contas ou, a valorização do tipo de conta seja alterada pelo administrador da plataforma.

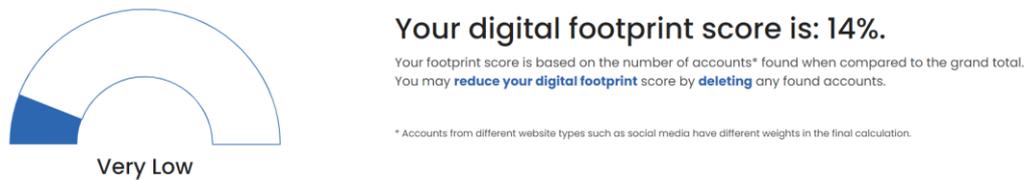


Figura 4.21 – Exemplo de classificação da pegada digital

A escolha encontra-se então do lado do utilizador. Caso opte pela remoção da conta via interface surgirá um *pop-up* com um guia para o mesmo (Figura 4.22).

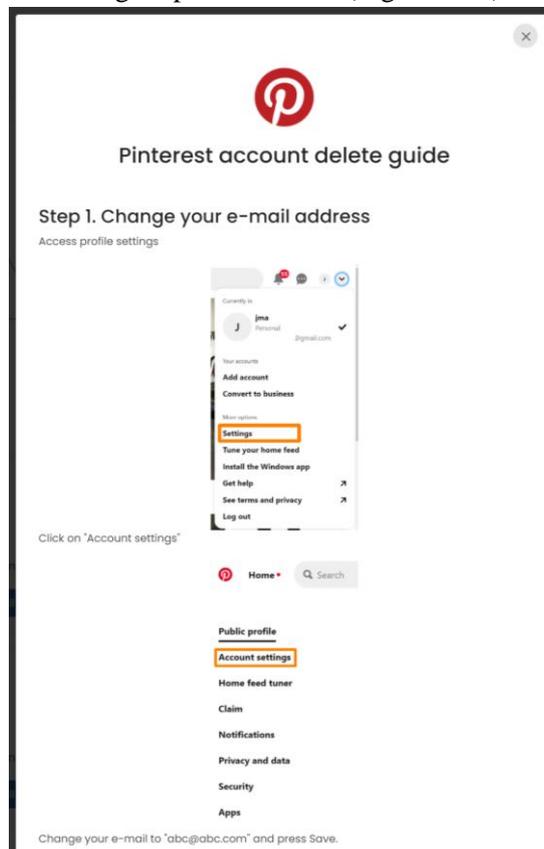


Figura 4.22 - Guia para a remoção da conta Pinterest via interface do utilizador

De outro modo, caso seja escolhida a opção do envio de *e-mail*, surge um *pop-up* com o texto modelo. Cabe agora ao utilizador optar por enviar ele próprio este *e-mail* ou permitir à aplicação que envie o mesmo em seu nome, preenchendo informação opcional (Figura 4.23).

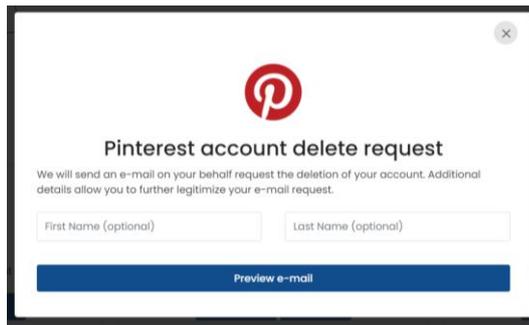


Figura 4.23 - Pedido de remoção via e-mail

Após definidos os campos opcionais será apresentada uma proposta de *e-mail* (Figura 4.24).

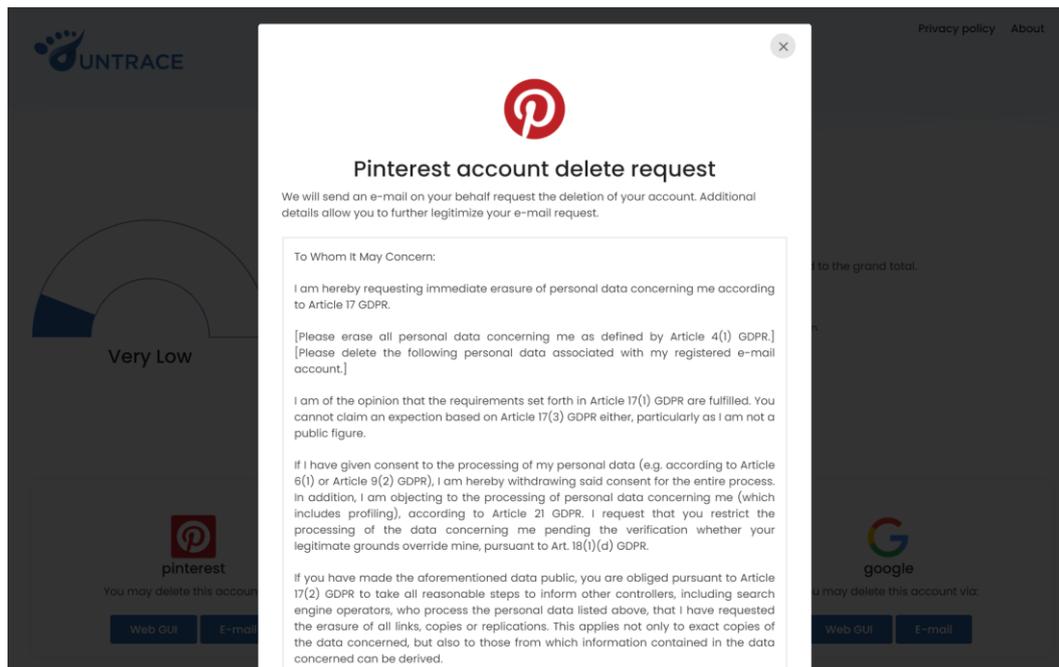


Figura 4.24 - Proposta de pedido via e-mail

Ao confirmar os dados (Figura 4.25) a solução Untrace irá utilizar novamente o serviço Sendgrid para o envio do *e-mail*, colocando o utilizador em CC para que este possa posteriormente acompanhar o pedido.

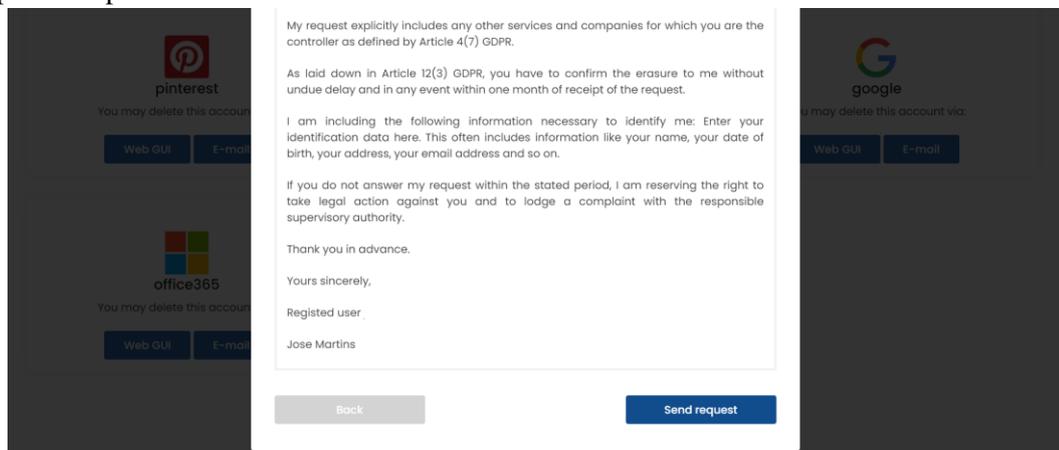


Figura 4.25 - Formulário de confirmação

Gestão da sessão do utilizador

Após efetuada uma pesquisa a sessão do utilizador tem uma duração de quinze minutos. Para renovar esta duração é iniciado um temporizador no navegador do utilizador (Figura 4.26).

```
function refreshsession($method, $url, $data){
  const res = await fetch('https://api.untraceapp.com/refreshToken', {
    method: 'POST',
    headers: {
      'Authorization': `Bearer ${store.JWT}`
    }
  });

  errorCatch(res);
  const jwt = await res.text();
  store.setJWT(jwt);
}

setTimeout(refreshsession(), 840000);
```

Figura 4.26 - Excerto de código JavaScript

Após catorze minutos é iniciado o pedido ao *endpoint* “refreshToken” para pedido de nova *token* de sessão seguida de uma validação do pedido. A função “*errorCatch*” irá validar a resposta do pedido e reagir conforme apropriado:

- *Token* inválida: inicia um pedido de criação de nova sessão;
- Erro na comunicação: despoleta um erro na consola;
- *Token* válida: o código é executado conforme esperado e a nova *token* é guardada para futuros pedidos.

No *backend* será feita uma comparação da *token* recebida no *header* “*Authorization*” com a *token* atualmente existente na base de dados correspondente ao ID da pesquisa. No caso da falha na comparação será devolvida a resposta apropriada e guardado o valor “*expired = 1*” na base de dados. Este valor será utilizado mais tarde no processo da gestão da informação.

Todas as funcionalidades descritas no decorrer deste capítulo foram implementadas e testadas internamente com sucesso, previamente à sua avaliação. A descrição das funcionalidades de Untrace servirão de base para o Capítulo 5 por forma a discutir as vantagens e desvantagens da mesma.

Capítulo 5

Avaliação da solução

A avaliação da Untrace teve por base uma análise quantitativa através do número de contas *online* encontradas bem como uma análise qualitativa, através de um questionário sobre a utilização da mesma. De entre as soluções existentes e discutidas no Capítulo 2 concluiu-se que apenas Mine cumpre com os objetivos de pesquisa e gestão da pegada digital. Por esta razão a análise quantitativa apenas compara os resultados de Mine e Untrace.

Enquanto consultora e prestadora de serviços de segurança de informação a Willway atua como fornecedora da Cybersecurity Cloud sobre o seu produto de sensibilização para a temática do *phishing*. Nesse sentido, a Untrace foi utilizada como complemento ao produto na fase recolha de informação sobre os utilizadores-alvo previamente ao lançamento de um simulacro de *phishing*. Nesta análise foi utilizada uma listagem de 20 endereços de *e-mail* do cliente “CYBER” (nome omitido).

A solução foi também avaliada sobre um contexto pessoal onde foi cedido um uso livre da Untrace para a consulta e gestão da pegada digital.

5.1.1 Avaliação quantitativa

Esta avaliação teve como objetivo medir os resultados do processo de pesquisa da pegada digital de Untrace e Mine. Os dados recolhidos referem-se ao número de contas *online* descobertas e à velocidade do processo de pesquisa de forma singular, com pedidos sequenciais, e paralela, com 20 pedidos de pesquisa em simultâneo.

Para a recolha destas métricas foi utilizado o navegador Google Chrome (v. 92.0.4515.15) e observada a janela das transferências da rede.

Tabela 5.1 – Métricas recolhidas

	# Resultados	Tempo médio de pesquisa singular	Tempo de pesquisa paralela
Untrace	40	~ 30s por endereço	~ 30s (sem diferença)
Mine	>80	25s a 2 mins por endereço	25s a 2 mins (sem diferença)

Untrace validou o registo dos 20 endereços de *e-mail* nos 121 *websites* presentes na base de dados de conhecimento (Tabela 4.4), tendo descoberto um total de 40 contas *online* (Tabela 5.1). A pesquisa sequencial sobre os 20 endereços levou um total de aproximadamente 10 minutos sendo que a pesquisa em paralelo encurtou esta tarefa para o total de 30 segundos. A paralelização dos pedidos não teve impacto no processo de pesquisa.

O processo de pesquisa de Mine envolveu a leitura das caixas de entrada dos 20 *e-mails* utilizados em busca de palavras-chave nos assuntos de *e-mail* e remetentes dos mesmos [67].

Através da correspondência de palavras-chave como “Amazon” e “LinkedIn” esta solução identifica as plataformas *web* nas quais o endereço de *e-mail* do utilizador se encontra registado. Observou-se um total de 80 registos de contas *online* entre os 20 endereços de *e-mail* e um tempo médio total para a pesquisa sequencial de 24 minutos, tendo a pesquisa em paralelo tomado um máximo de 2 minutos.

5.1.2 Avaliação qualitativa

Para a análise qualitativa 16 utilizadores foram convidados a usar a solução com o seu endereço de *e-mail* pessoal, seguida de um questionário sobre a mesma. O questionário teve como objetivo medir a perceção dos utilizadores sobre a sua pegada digital através da comparação do número de contas *online* aparentemente desconhecidas pelos mesmos face o total encontrado, perceber em que medida os mesmos tomaram uma ação sobre a sua pegada digital e identificar pontos de melhoria sobre a usabilidade da solução.

Questão: Quantos resultados obteve associados ao seu *e-mail* pessoal?

Tabela 5.2 - Resultados da pesquisa Untrace sobre o e-mail pessoal

Resposta possível	Resultado
< 10 contas encontradas	19% (3)
10 a 30 contas encontradas	63% (10)
30 a 50 contas encontradas	12% (2)
> 50 contas encontradas	6% (1)

Questão: Quantos dos resultados observados lhe eram desconhecidos?

Tabela 5.3 - Avaliação dos resultados obtidos

Resposta possível	Resultado
< 25%	87% (14)
25 a 50%	13% (2)
50 a 75%	0
> 75%	0

Questão: Iniciou o processo ou cancelou alguma das contas encontradas via interface? Se sim, quantas?

Tabela 5.4 - Número de contas canceladas ou em processo

Resposta possível	Resultado
< 25%	100% (16)
25 a 50%	0
50 a 75%	0
> 75%	0

Questão: Classifique a usabilidade da plataforma

Tabela 5.5 - Classificação de usabilidade da Untrace

Resposta possível	Resultado
Nada fácil de utilizar	0
Pouco fácil de utilizar	6% (1)
Algo fácil de utilizar	13% (2)
Muito fácil de utilizar	75% (12)
Extremamente fácil de utilizar	6% (1)

Questão: Classifique o seu nível de sensibilização prévio para o tema “Pegada Digital”

Tabela 5.6 - Avaliação do nível de sensibilização para o tema "pegada digital"

Resposta possível	Resultado
Nada familiar	25% (4)
Pouco familiar	62% (10)
Algo familiar	13% (2)
Muito familiar	0
Extremamente familiar	0

Questão: Oportunidades de melhoria / sugestões?

Tabela 5.7 - Sugestões de melhoria

Resposta aberta
Melhorias ao UI / UX

Através dos resultados obtidos observou-se que até um quarto das contas *online* encontradas pela Untrace eram desconhecidas pelo utilizador, a gestão da pegada digital resultou na redução de menos de 25% da presença *online* do mesmo, grande parte dos utilizadores (62%) estava pouco familiarizado com o tema e a maioria (75%) percecionou a solução como fácil de utilizar.

5.1.3 Discussão

Untrace cumpriu o objetivo definido pela Willway de auxiliar o processo de recolha de informação sobre os endereços de *e-mail* para a preparação de um simulacro de *phishing*. Foi possível observar que entre os 40 resultados todos os 20 endereços se encontravam registados na plataforma *web* Atlassian, utilizada posteriormente como tema da simulação.

Por forma a alcançar o mesmo objetivo utilizando Mine é necessário conceder um acesso de leitura sobre a caixa de entrada de cada um dos *e-mails* dos utilizadores. Esta autorização adiciona uma complexidade técnica e legal ao objetivo da Willway que invalida a utilização da solução e por conseguinte o incumprimento do objetivo.

Apesar disto, ambas as soluções foram utilizadas para uma comparação de resultados. Apesar de Mine ter identificado de forma sobreposta o dobro de contas *online* de Untrace (80 vs. 40), o tempo de pesquisa foi também mais demorado que o de Untrace (2 minutos vs. 30 segundos) pois é proporcional ao tamanho da caixa de entrada do *e-mail* do utilizador.

Em suma, foi possível obter uma maior identificação da pegada digital utilizando Mine. No entanto, este resultado é apenas possível de obter através da concessão de uma permissão crítica, ao contrário do que acontece no processo de pesquisa de Untrace onde não é necessário qualquer tipo de acesso. Ainda assim, esta comparação prolonga-se. Por não necessitar de qualquer acesso, Untrace pesquisa toda a *internet* por registos do endereço de *e-mail* baseando-se na base de dados de conhecimento da solução. Esta base de dados, por sua vez gerida de forma manual, nunca obterá os mesmos resultados de Mine, cujo mecanismo de pesquisa se baseia na comparação de palavras-chave predefinidas (ex. “Facebook”, “LinkedIn”) [67]. Untrace procura combater esta desvantagem agregando bases de dados públicas como as identificadas no Capítulo 2.

Através do questionário de utilização foi possível concluir que apesar de até um quarto das contas *online* encontradas serem desconhecidas pelo próprio utilizador tal não implica que o mesmo tome a ação de as eliminar (Tabela 5.3). Apesar de os utilizadores não terem reduzido a sua pegada digital de forma considerável, Untrace cumpre o objetivo definido de permitir a gestão da mesma através de uma plataforma fácil de utilizar (Tabela 5.5).

Tabela 5.8 - Comparação de funcionalidades entre Untrace e Mine

Funcionalidade	Solução	
	Untrace	Mine
Pesquisa por <i>e-mail</i>	✓	X
Pesquisa da <i>inbox</i>	X	✓
Avaliação da pegada digital	✓	✓
Remoção da conta de utilizador via interface	✓	X
Remoção da conta de utilizador via <i>e-mail</i>	✓	✓
Customização do <i>e-mail</i> formal	✓	X
Sensibilização sobre o tema	✓	✓

De forma a facilitar a comparação entre Untrace e Mine, foi construída a Tabela 5.8. Através da mesma é ainda possível destacar que Untrace oferece guias gráficos para a remoção da conta de utilizador, mecanismo que procura combater a dificuldade inerente à gestão da reputação no envio de *e-mails* [68]. Untrace possibilita também a customização do *e-mail* de pedido de remoção de registo.

Capítulo 6

Conclusões e trabalho futuro

6.1 Conclusões

Este trabalho teve como objetivos apresentar uma solução que facilitasse a gestão da pegada digital do utilizador-final complementando as atividades de auditoria e sensibilização para o *phishing* enquanto serviço prestado pela Willway.

Ao dia de hoje existem algumas ferramentas que tentam cumprir este propósito, no entanto, são intrusivas e requerem o acesso à toda a caixa de correio do utilizador por forma a analisar em quais os serviços ou plataformas *web* em que este se encontra registado. A exceção é Holehe, uma ferramenta de código-livre que procura o registo do *e-mail* do utilizador sem necessitar de qualquer outro tipo de permissão.

Através de um questionário de avaliação de necessidade foi possível entender que os utilizadores estão pouco sensibilizados para a temática da “pegada digital”, sendo que lhes seria útil uma plataforma através da qual fosse possível descobrir em que lojas, fóruns e ou redes sociais o seu *e-mail* poderia encontrar-se registado.

O desenvolvimento da solução teve como objetivo a criação de uma plataforma digital através da qual os seus utilizadores pudessem encontrar e\ou remover contas associadas ao seu endereço de *e-mail* com o objetivo final de gerir a pegada digital. A solução foi desenvolvida utilizando tecnologias que abordam as questões de paralelismo computacional e integração com futuras soluções. A utilização encontra-se de momento restrita a clientes e projetos da Willway.

Como forma de avaliação, a solução esta foi utilizada num projeto de sensibilização para o *phishing*. De entre 20 *e-mails* recolhidos foi possível observar que os mesmos partilhavam um registo na Atlassian, utilizada posteriormente como tema de uma simulação de *phishing*. No fim os utilizadores foram convidados a utilizar a plataforma para uso pessoal. Na comparação entre Untrace e a solução concorrente Mine foi possível concluir que a Untrace não é intrusiva pois não requer qualquer tipo de acesso por parte do utilizador e disponibiliza um mecanismo adicional para a remoção da conta do utilizador, através de um guia de interface.

6.2 Trabalho futuro

O processo de pesquisa da solução apresenta dificuldades na gestão da base de dados conhecimento. Atualmente esta é feita de forma manual embora se procure integrar outras fontes de informação públicas como as identificadas no Capítulo 2. A automatização desta gestão levanta a necessidade de criar um mapa de plataformas *web*, útil para identificar onde consultar a pegada digital e ainda um algoritmo de validação de registo, útil no processo de como consultar a pegada digital. Como sugestão para a obtenção de um mapa de plataformas *web* poderia ser desenvolvido um trabalho de análise às entidades de registo DNS. Por outro lado, para automatizar o processo de validação de registo de um endereço de *e-mail*, em nota de trabalho futuro, poder-se-ia treinar

algoritmos de inteligência artificial por modo a identificar *endpoints* de registo de conta e criação de pedidos sobre os mesmos.

A expansão desta base de dados implicaria por sua vez uma revisão do processo de pesquisa garantindo que a paralelização é feita de forma eficiente. Existem atualmente *frameworks web* baseadas em Python que procuram obter velocidades de paralelização superiores [69]. Estas poderão ser consideradas em conjunto com uma reformulação da infraestrutura por modo a suportar o processamento de uma forma distribuída por diferentes servidores tolerantes a faltas.

Numa perspetiva processual, o âmbito da utilização da solução poderá ser expandido conforme a estratégia de negócio da Willway. A solução encontra-se pronta para ser desenvolvida a uma maior escala e até mesmo ser tornada pública.

A sensibilização para a pegada digital é certamente um tema que irá crescer a par com a *internet* e os serviços por esta disponibilizados. Deste modo, a solução poderá beneficiar de uma comunidade *online* interessada na privacidade do utilizador. A abertura à colaboração sobre o código-fonte poderá ainda trazer proveitos como o aumento de *websites* disponíveis para pesquisa e melhorias ao processo de remoção de contas de utilizador.

Anexo I – Política de privacidade

Controller

Willway, S.A. with headquarters in Avenida da República, 50, 2º, 1050-196 Lisboa, taxpayer number 509713335.

For questions related to the processing of your personal data, including the exercise of your rights, you should contact us via email to privacy@willway.pt.

Purposes for which we process your data when collected through this site

- **User account search**

If you introduce your *e-mail*, Willway may keep all related user account search results until your user session is over or such data is expired. All user-related result data is expired within 16 minutes and deleted daily. Management of job applications

- **Cookies**

This site includes cookies used for maintaining your search results active.

Your Rights

Under certain circumstances, you have the rights to:

Access, correct, erasure, object to processing, restrict the processing, portability of your personal data;

Withdraw, at any moment, consents that you have given, which does not invalidate the data processing operations that were carried out in the period when the consent existed;

Lodge a complaint to Supervisory Authority. In Portugal, the Supervisory Authority is CNPD (Comissão Nacional de Proteção de Dados).

Entirely automated decisions, including profiling

You will not be subject to decisions that will have a significant impact on you based solely on automated decision-making.

Bibliografia

- [1] CISCO, “Cisco Annual Internet Report (2018–2023) White Paper,” 3 03 2020. [Online]. Available: <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>.
- [2] M. M. a. P. J. Michael E. Locasto, “Security and privacy considerations in digital death,” *Association for Computing Machinery*, n° In Proceedings of the 2011 New Security Paradigms Workshop (NSPW '11), p. 1–10, 2011.
- [3] Autoridade Tributária, [Online]. Available: <https://info.portaldasfinancas.gov.pt/pt/destaques/Paginas/default.aspx>. [Acedido em 07 09 2021].
- [4] CTT, [Online]. Available: <https://www.ctt.pt/transversais/alertas-de-phishing>. [Acedido em 07 09 2021].
- [5] M. D. d. J. Susanne Barth, “The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review,” *Telematics and Informatics*, vol. 34, n° 7, pp. 1038-1058, 2017.
- [6] L. P. a. P. Dourish, “Unpacking "privacy" for a networked world,” *Association for Computing Machinery*, n° SIGCHI Conference on Human Factors in Computing Systems, p. 129–136, 2003.
- [7] C. C. M. R. B. A. S. a. T. R. Siân E. Lindley, “Rethinking the web as a personal archive,” *Association for Computing Machinery*, n° In Proceedings of the 22nd international conference on World Wide Web (WWW '13), p. 749–760, 2013.
- [8] Glassdoor, “Delete account help,” [Online]. Available: https://help.glassdoor.com/s/article/Close-my-Glassdoor-account?language=en_US#:~:text=Click%20the%20Glassdoor%20Profile%20icon,Click%20Close%20Account..
- [9] Battle.net, “Delete account help,” [Online]. Available: <https://us.battle.net/support/en/article/2659>. [Acedido em 01 12 2021].
- [10] Alibaba, “Delete Account Help,” 01 10 2021. [Online]. Available: <https://service.alibaba.com/page/search?pageId=125&keyword=delete&language=en>.
- [11] OWASP, “OWASP - Information Gathering,” 03 12 2020. [Online]. Available: https://owasp.org/www-project-web-security-testing-guide/stable/4-Web_Application_Security_Testing/01-Information_Gathering/. [Acedido em 23 08 2021].
- [12] Priberam, “Definição de pegada,” 03 04 2021. [Online]. Available: <https://dicionario.priberam.org/pegada>.
- [13] Priberam, “Definição de digital,” 04 03 2021. [Online]. Available: <https://dicionario.priberam.org/digital>.
- [14] F. C. F. D. F. C. R. a. J. B. F. Girardin, “Digital Footprinting: Uncovering Tourists with User-Generated Content,” *Pervasive Computing, IEEE 7.4*, pp. 36-43, 2008.
- [15] Britannica, “Telemetry,” 09 07 2021. [Online]. Available: <https://www.britannica.com/technology/telemetry>. [Acedido em 07 09 2021].
- [16] Newwegg, [Online]. Available: <https://www.newweggbusiness.com/smartbuyer/windows/should-you-disable-windows-10-telemetry/>. [Acedido em 07 09 2021].

- [17] Fosspost. [Online]. Available: <https://fosspost.org/audacity-is-now-a-spyware/>. [Acedido em 07 09 2021].
- [18] C. & H. C. Lutz, "The Dark Side of Online Participation: Exploring Non-, Passive and Negative Participation," *nformation Communication and Society* 20, pp. 876-897, 2017.
- [19] A. L. Y. & A. Quan-Haase, "Privacy Protection Strategies on Facebook," *Information, Communication & Society*, pp. 479-500, 2013.
- [20] L. E. K. A. Azu MC, "Social Media, Surgeons, and the Internet: An Era or an Error?," *The American Surgeon*, pp. 555-558, 2012.
- [21] H. B. e. A. a. K. N. B. NARGESIAN, "Digital footprint in Web 3.0: Social Media Usage in Recruitment," 2019. [Online]. Available: Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-02792019000100139&lng=en&nrm=iso. [Acedido em 07 10 2021].
- [22] "OSINT Framework," [Online]. Available: <https://osintframework.com>. [Acedido em 27 07 2021].
- [23] Skymem. [Online]. Available: <https://www.skymem.info/>. [Acedido em 07 09 2021].
- [24] Mailshunt. [Online]. Available: <https://www.mailshunt.com/>. [Acedido em 07 09 2021].
- [25] Hunter.io. [Online]. Available: <https://hunter.io/>. [Acedido em 07 09 2021].
- [26] "Hunter.io - Our Data," [Online]. Available: <https://hunter.io/our-data>. [Acedido em 18 09 2021].
- [27] Infoga. [Online]. Available: <https://github.com/m4ll0k/infoga>. [Acedido em 07 09 2021].
- [28] TheHarvester. [Online]. Available: <https://github.com/laramies/theHarvester>. [Acedido em 07 09 2021].
- [29] haveIbeenPwned. [Online]. Available: <https://haveibeenpwned.com/>. [Acedido em 07 09 2021].
- [30] Dehashed. [Online]. Available: <https://dehashed.com/>. [Acedido em 07 09 2021].
- [31] Unroll.me, "Privacy Policy," [Online]. Available: <https://unroll.me/legal/privacy>. [Acedido em 01 12 2021].
- [32] Mine, "Privacy Policy," [Online]. Available: <https://saymine.com/privacy-policy/the-highlights>. [Acedido em 01 12 2021].
- [33] Mine. [Online]. Available: <https://saymineapp.com>. [Acedido em 07 09 2021].
- [34] Usersearch. [Online]. Available: <https://www.usersearch.org/>. [Acedido em 01 08 2021].
- [35] Unroll.Me Inc., [Online]. Available: <https://unroll.me/>. [Acedido em 15 07 2021].
- [36] Megadose. [Online]. Available: <https://github.com/megadose/holehe>. [Acedido em 07 09 2021].
- [37] Background Checks.org, [Online]. Available: <https://backgroundchecks.org/justdeleteme/>. [Acedido em 14 07 2021].
- [38] ACCOUNTKILLER.COM.
- [39] Abine, Inc, [Online]. Available: <https://joindeleteme.com/sites-we-remove-from/>. [Acedido em 14 07 2021].
- [40] J. & C. J. & L. T. Blue, "Digital Footprints: Your Unique Identity," *British HCI* 2018, pp. 1-5, 2018.
- [41] D. Piper. [Online]. Available: <https://www.dlapiperdataprotection.com/index.html?t=world-map>. [Acedido em 18 09 2021].
- [42] SecureCoding. [Online]. Available: <https://www.securecoding.com/blog/penetration-testing-steps/>. [Acedido em 18 09 2021].
- [43] Hasura.io, "Silent Refresh JWT," [Online]. Available: https://hasura.io/blog/best-practices-of-using-jwt-with-graphql/#refresh_token. [Acedido em 19 09 2021].

- [44] Smartbug. [Online]. Available: <https://www.smartbugmedia.com/blog/update-website>. [Acedido em 20 09 2021].
- [45] National Institute of Standards and Technology, “Digital Identity Guidelines,” 2017.
- [46] LetsEncrypt, “Let’s Encrypt,” 21 09 2021. [Online]. Available: <https://letsencrypt.org>.
- [47] StackOverflow, “Most popular web frameworks 2021,” [Online]. Available: <https://insights.stackoverflow.com/survey/2021#section-most-popular-technologies-web-frameworks>. [Acedido em 20 12 2021].
- [48] A. C. a. C. B. W. Tobias Lauinger, “Thou shalt not depend on me,” *ACM*, p. 41–47, 2018.
- [49] I. & B. S. & A. A. & S. P. Zozas, “Estimating the Maintenance Effort of JavaScript Applications,” *IEEE*, vol. 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
- [50] React, [Online]. Available: <https://reactjs.org/>. [Acedido em 02 12 2021].
- [51] jQuery, [Online]. Available: <https://jquery.com/>. [Acedido em 02 12 2021].
- [52] jQuery, “API Documentation,” [Online]. Available: <https://api.jquery.com/>. [Acedido em 02 12 2021].
- [53] Mathias. [Online]. Available: <https://mathiasbynens.be/demo/jquery-size>. [Acedido em 01 12 2021].
- [54] jQuery, [Online]. Available: <https://jquery.com/browser-support>. [Acedido em 02 12 2021].
- [55] Gartasio, [Online]. Available: <https://blog.gartasio.com/you-dont-need-jquery/selectors>. [Acedido em 02 12 2021].
- [56] Express, [Online]. Available: <https://expressjs.com/>. [Acedido em 03 12 2021].
- [57] Mozilla, “Express/Node introduction,” [Online]. Available: https://developer.mozilla.org/en-US/docs/Learn/Server-side/Express_Nodejs/Introduction. [Acedido em 03 12 2021].
- [58] Angular, [Online]. Available: <https://angular.io/>. [Acedido em 03 12 2021].
- [59] Vuejs, [Online]. Available: <https://vuejs.org/>. [Acedido em 03 12 2021].
- [60] L. E. Hecht, “This Week in Numbers: React Leads the Pack, Though Vue.js Makes Big Strides in China,” [Online]. Available: <https://thenewstack.io/vue-js-big-china/>. [Acedido em 03 12 2021].
- [61] Quart, [Online]. Available: <https://pypi.org/project/Quart/>. [Acedido em 21 09 2021].
- [62] Python, [Online]. Available: <https://docs.python.org/3/library/asyncio.html>. [Acedido em 04 12 2021].
- [63] “Flask,” [Online]. Available: <https://flask.palletsprojects.com/en/2.0.x/>. [Acedido em 21 09 2021].
- [64] Python, “Coroutines and Tasks,” [Online]. Available: <https://docs.python.org/3/library/asyncio-task.html>. [Acedido em 23 10 2021].
- [65] MongoDB, “MongoDB vs MySQL,” [Online]. Available: <https://www.mongodb.com/compare/mongodb-mysql>. [Acedido em 21 09 2021].
- [66] Sendgrid. [Online]. Available: <https://sendgrid.com/>. [Acedido em 22 09 2021].
- [67] Mine, “About us,” [Online]. Available: <https://saymine.com/about>. [Acedido em 07 09 2021].
- [68] B. Jackson, “Email Authentication – Don’t Let Your Emails End Up in Spam,” [Online]. Available: <https://kinsta.com/blog/email-authentication/>. [Acedido em 03 12 2021].
- [69] K. Klenov, “Python Async (ASGI) Web Frameworks Benchmark,” [Online]. Available: <http://klen.github.io/py-frameworks-bench/>. [Acedido em 15 11 2021].