DEPARTAMENTO DE ENGENHARIAS E CIÊNCIAS DA COMPUTAÇÃO

MESTRADO EM ENGENHARIA INFORMÁTICA E DE TELECOMUNICAÇÕES

UNIVERSIDADE AUTÓNOMA DE LISBOA

"LUÍS DE CAMÕES"

CONSTRUÇÃO DE LINKED DATA MASHUP PARA A INTEGRAÇÃO DE DADOS DE ACESSO AO ENSINO SUPERIOR PORTUGUÊS

Dissertação para a obtenção do grau de Mestre em Engenharia Informática e de Telecomunicações

Autor: Luciano Soares Martins

Orientadora: Professora Doutora Valéria Magalhães Pequeno

Número do candidato: 30006009

Junho de 2021

Lisboa

Dedicatória

Dedico esse trabalho aos meus mentores do mundo alternativo que apesar de ausentes e inexistentes estão ali para indicar que há sempre um desafio iminente e insólito.

Agradecimentos

Correndo o risco de esquecer alguém importante, agradeço a minha família e todos professores e colegas da Universidade Autónoma de Lisboa que de alguma forma contribuíram para o desenvolvimento deste trabalho, entretanto agradeço em especial ao Luis Pedroso, cuja contribuição na tradução para o português de Camões, foi de muita importância.

Resumo

O presente trabalho parte do princípio de que a escolha de um curso superior requer ampla pesquisa numa grande variedade de páginas na *web* de cursos oferecidos por inúmeras universidades. Os cursos disponíveis, mesmo com nomes iguais podem ser totalmente diferentes de uma instituição para outra. Além disto, as saídas profissionais podem não estar de acordo com o esperado pelo utilizador tendo em conta sua expectativa profissional e de futuro.

Neste cenário, a criação de uma ferramenta que consolide essas informações num único local de modo a facilitar a tomada de decisão apresenta-se como uma medida viável. Imagina-se que seja possível reunir essas informações vindas de bases de dados externas ou mesmo *websites* e sem formato estruturado mantendo a integridade e atualidade dos dados. O foco é apresentar essas informações integradas no formato de dados abertos que permitam o acesso não só por pessoas mas também por máquinas. Para isso optou-se pelo uso de tecnologias da *web* semântica, de modo a permitir de forma mais específica a integração semântica dos dados disponíveis na *web* referentes aos cursos oferecidos pelas instituições do ensino superior e os dados relativos às saídas profissionais, materializados na forma de ofertas de emprego disponíveis em *websites* especializados.

A proposta do presente trabalho é a criação de um portal que possa demonstrar os processos de pesquisa de informações em vários sítios e formatos na *web*, fazer a integração dessas informações utilizando ferramentas da *web* semântica e, por fim, a apresentação em formato aberto e estruturado das informações integradas. O portal é direcionado ao utilizador técnico que pretende usar as ferramentas produzidas como base para a criação de um outro portal voltado ao utilizador final, a saber os alunos prestes a escolher um curso superior para início da carreira profissional.

Palavras-chave: *web* semântica; RDF; RDFS, OWL; *linked data mashup*; ontologia; carreira profissional; curso superior; saída profissional, dados abertos.

Abstract

The present work assumes that choosing a higher education course requires extensive research in a wide variety of web pages for courses offered by numerous universities. The available courses, even with the same names, can be totally different from one institution to another. In addition, professional outings may not be as expected by the user, taking into account their professional and future expectations.

In this scenario, the creation of a tool that consolidates this information in a single place in order to facilitate decision making is a viable measure. It is thought that it is possible to gather this information coming from external databases or even websites and without a structured format, maintaining the integrity and timeliness of the data. The focus is to present this information integrated in an open data format that allows access not only by people but also by machines. For this, we opted for the use of semantic web technologies, in order to allow more specifically the semantic integration of data available on the web relating to courses offered by higher education institutions and data relating to professional opportunities, materialized in the form of job offers available on specialized websites.

The purpose of this work is to create a portal that can demonstrate the information search processes in various sites and formats on the web, integrate this information using semantic web tools and, finally, the presentation in an open and structured format. of the integrated information. The portal is aimed at the technical user who intends to use the tools produced as a basis for creating another portal aimed at the end user, namely students about to choose a higher education course to start their professional career.

Keywords: semantic web; RDF; RDFS, OWL; linked data mashup; ontology; professional career; higher education; professional output, open data.

Índice

Lista de Tabelas

Lista de Figuras

1 Introdução

O processo de tomada de decisão de seleção da carreira profissional é um dos elementos-chave na vida de um indivíduo. A escolha certa do curso superior para os alunos é crítico tendo alto impacto na sua vida profissional e realizações futuras. Este é um assunto que não pode ser deixado na intuição de cada um, em noções preconcebidas, em imaginações ou em conceitos populares.

A ausência de escolha ou escolha não acertada pode direcionar todos esforços individuais num caminho errado, quando não alinhada com as expectativas. De qualquer forma, o realinhamento posterior é possível, porém ele não seria apenas frustrante, mas um desperdício de recursos, tempo e motivação.

A escolha da carreira profissional dos alunos precisa ser baseada em forte conhecimento, informações completas e apropriadamente guiadas, combinando personalidade individual, tipo de carreira profissional e outros fatores intrínsecos e extrínsecos, explicados mais a frente. Os alunos precisam ser orientados em novas tendências emergentes, oportunidades futuras e desafios no contexto de opções de escolha de carreira. Eles precisam conhecer as tendências predominantes do mercado, práticas e cenário de trabalho de vários setores.

No âmbito desse projeto de mestrado, propõe-se a recolha, armazenamento, integração e disponibilização de informação de forma facilitada para a análise de opções viáveis para escolha de faculdade/universidade e curso, enquanto ferramenta de tomada de decisão. O caminho escolhido foi a construção de um *linked data mashup* (LDM) para a integração de dados de acesso ao ensino superior português. A construção de qualquer LDM pressupõe a busca de dados em formatos estruturados e/ou não estruturados para criar uma visão homogénea que, materializada, pode permitir o desenvolvimento de aplicações como listas, tabelas, gráficos, *dashboard*, etc., e fornecer informações consolidadas.

O presente projeto recolheu dados de cursos em *websites* de instituições de ensino superior, organizou de forma estruturada e criou uma referência cruzada com as saídas profissionais, materializadas em ofertas de trabalho através de *websites* de busca de emprego. No limite, como exemplo, utilizou dados de uma única instituição e consolidou informações sobre o mercado de trabalho de diversas áreas, oriundos de um *website* de ofertas de emprego.

Pretende-se com o presente trabalho constituir uma base de pesquisa que permita a exploração das ferramentas da *web* semântica no ramo académico de modo a auxiliar utilizadores de nível técnico a prover informações aos utilizadores finais, haja vista que as

iniciativas de utilização da *web* semântica no mundo acadêmico com este formato de recolha de dados de diversas fontes, armazenamento e integração, têm sido pouco utilizadas quando comparadas a outras iniciativas na *internet*.

1.1 Problema

Apesar do crescente volume de dados disponíveis na *web*, informações relevantes sobre um mesmo curso podem encontrar-se em bases de dados e locais diferentes, disponibilizadas em formatos proprietários, como folhas de cálculo e *websites*, impossibilitando a integração e consulta dessas informações de maneira simplificada por terceiros. Posto isto, algumas questões relevantes se apresentam:

1 É possível reunir as informações necessárias vindas de bases de dados externas ou mesmo *websites* e sem formato estruturado mantendo a sua integridade e atualidade dos dados de modo que se possa apresentar esses mesmos dados de forma aberta e estruturada?

2 Pode o uso de tecnologias da w*eb semântica* como o LDM ser alternativa para superar os problemas de recolha dessas informações, integrando semanticamente os dados e gerando como resultado uma vista homogénea?

Na hipótese das respostas às duas perguntas serem positivas, considera-se que é possível criar um portal com as informações recolhidas de forma estruturada e homogénea, que permita facilmente a tomada de decisão sobre qual a melhor carreira profissional a seguir.

1.2 Objetivos

O objetivo principal deste trabalho é a criação de um LDM para integração semântica dos dados disponíveis na *web* referentes aos cursos oferecidos pelas instituições do ensino superior e os dados relativos às saídas profissionais, materializados na forma de ofertas de emprego disponíveis em *websites* especializados para o território português.

Para tanto, as tarefas seguintes são partes imprescindíveis para se atingir este objetivo:

1) Reuso ou criação de uma ontologia para estruturar e guardar os dados sobre o cenário em estudo;
2) Extração, transformação e consolidação dos dados disponíveis na *web* que sejam relevantes para a tomada de decisão sobre o acesso ao ensino superior português;
3) Criação de um LDM para a integração semântica dos dados relevantes sobre o acesso ao ensino superior português;

4) Disponibilização desses dados em formato aberto para possível uso por terceiros e para o desenvolvimento de aplicações num portal disponibilizado na *web.*

Com isto, será possível a partir deste trabalho, gerar informações, ainda que numa interface mais técnica, que permita ao utilizador final a facilidade na tomada de decisões sobre o futuro profissional de forma mais assertiva e eficaz.

Como objetivo secundário, pretende-se que este trabalho seja base de estudo para a criação de ferramentas de pesquisa com interface mais amigável e intuitiva a partir de dados abertos.

1.3 Justificação

Escolher um curso numa universidade não é tarefa fácil para os alunos que, em determinada altura, têm que tomar esta decisão. Uma grande variedade de cursos é oferecido por universidades e instituições, cuja informação e requisitos de acesso são totalmente díspares entre elas.

Encontrar informações relevantes sobre o ensino superior a partir de um grande número de *websites* é um processo desafiador e demorado e percorrer a infinidade de cursos disponíveis para atender às suas necessidades individuais é uma experiência exaustiva.

Essas informações abundantes significam que os alunos precisam pesquisar, organizar e usar os recursos que podem permitir que eles correspondam aos seus objetivos individuais, interesses e nível atual de conhecimento de forma adequada. Isso pode ser um processo demorado, pois envolve o acesso a cada plataforma, procurando os cursos disponíveis, leitura e análise atenta ao programa de cada curso e, em seguida, a escolha daquele que é mais adequado.

No entanto, mesmo que as informações fornecidas pelas universidades nos seus *websites* estejam cada vez mais claros, não significa automaticamente que os alunos possuem a capacidade cognitiva de avaliar cada um dos cursos. Em vez disso, eles são confrontados com um problema denominado “sobrecarga de informação”.

Pode acabar por ser frustrante, depois de exaustivo trabalho de pesquisa e tantos dados analisados, não conseguir ainda responder às questões que se colocam:

- Qual faculdade/universidade escolher?
- Qual curso fazer?
- Qual profissão pretendida após a faculdade?

Responder a essas questões é algo que terá um grande impacto não somente durante os três a cinco anos dedicados ao curso superior, mas também durante a vida profissional do aluno. Esse tempo de estudos é ainda mais extenso quando se trata de cursos de medicina e afins que exigem mais alguns anos de estágio e/ou residência.

A maior parte das pessoas pode ficar numa encruzilhada e acaba por tomar a decisão por impulso, de forma emocional e esta não é a melhor forma de decidir algo tão importante. É necessário um método ou um sistema testado e aprovado que garanta ao máximo a possibilidade de não ser escolhida a melhor opção disponível.

Instintivamente os alunos costumam seguir os seguintes passos, valorizando e enfatizando uns a mais e outros menos:

1) Visualizar o seu futuro;
2) Reduzir as opções a uma das três grandes áreas;
3) Analisar as opções viáveis dentro da área escolhida;
4) Analisar tendências de mercado;
5) Tomar a decisão.

A despeito do valor que cada aluno concede para esses passos, o item 3, que é analisar as opções viáveis dentro da área escolhida, é inevitavelmente um passo que recebe maior foco e atenção, logo, em função disto, o corrente trabalho está focado nesse passo. Assim sendo, serão tratadas, dentre outras, as seguintes informações:

1) Faculdades/universidades existentes;
2) Áreas de conhecimento;
3) Cursos superiores disponíveis;
4) Modalidades de ensino;
5) Saídas profissionais.

Pelo que se pode observar até o momento, não há uma base simples de consulta. As informações encontram-se dispersas pela *web* sem estruturação dos dados, o que dificulta a recolha e análise das mesmas e força o utilizador a dispensar esforço em pesquisas não automatizadas, cuja assertividade deixa a desejar, trazendo mais riscos de uma decisão ser tomada sem todas as informações disponíveis.

A *web* de dados tem-se mostrado em grande crescimento nos últimos anos. Uma das suas características é que se baseia no *Linked Data* (LD), que é um conjunto de melhores práticas para publicação e utilização de dados estruturados, o que permite a integração de dados

entre diferentes bases de dados numa representação uniforme e num único espaço de dados global [76]. As estruturas de suporte do LD estão nas tecnologias da *web* semântica e permitem reduzir a complexidade de integrações por causa das ligações estabelecidas e descritas entre os conjuntos de dados. Para além do referido, o uso de um modelo de dados padronizado e um mecanismo de consulta também padronizado, por exemplo SPARQL[1], simplificam mais ainda a integração dos dados.

Ao utilizar a *web* semântica como extensão da *web*, incluindo as suas ferramentas e protocolos, espera-se atingir os objetivos propostos e gerar valor ao utilizador final que, de forma dinâmica e eficaz, tenha condições de tomar decisões sobre o futuro profissional sem esforço técnico de pesquisa e análise preditiva, podendo concentrar a energia e tempo em informações consolidadas de qualidade.

1.4 Limitações

Não se pretende com este estudo, resolver todos os problemas para todos os alunos. De modo a alcançar resultados concretos, foram definidos limites e critérios para o desenvolvimento deste trabalho.

O primeiro limite é que serão considerados somente cursos de licenciatura ou licenciatura e mestrados integrados, de universidades e institutos portugueses.

Para isto, optou-se numa primeira instância pela recolha de dados em fontes de informação públicas focadas no tema da educação, a saber, o "Portal dos dados abertos da administração pública"[2] e o "Portal da Orientação Vocacional"[3]. Como neste último, os dados são de médias, exames e outras informações relacionadas, que fogem ao âmbito deste trabalho, foi utilizado como ponto de partida as informações relacionadas com as universidades e institutos portugueses disponíveis do "Portal dos dados abertos da administração pública".

A tabela disponível no formato JSON[4] disponível no endereço[5] no referido portal contém as informações de todas instituições de ensino superior portuguesas. A partir desta lista, para a construção do artefacto no âmbito deste trabalho, foram escolhidas as 10 melhores de acordo com o *ranking* do *website* "*The Times Higher Education World University Rankings*"[6].

---

[1] https://en.wikipedia.org/wiki/SPARQL [Consult. em 23/08/2021].
[2] https://dados.gov.pt/pt/ [Consult. em 23/08/2021].
[3] https://ov.portalpsi.net/medias/ [Consult. em 23/08/2021].
[4] https://pt.wikipedia.org/wiki/JSON [Consult. em 23/08/2021].
[5] https://dados.gov.pt/pt/datasets/r/01db2c07-0739-4cde-a481-6fe6aed34b01 [Consult. em 23/08/2021].
[6] https://www.timeshighereducation.com [Consult. em 23/08/2021].

A escolha desse *website* para classificar a universidades portuguesas foi principalmente pela análise de sua metodologia, considerada a mais apropriada, pois avalia as universidades em todas as suas missões principais: ensino, pesquisa, transferência de conhecimento e perspetiva internacional. O *website* usa treze indicadores de desempenho cuidadosamente calibrados para fornecer as comparações mais abrangentes e equilibradas, com a confiança de estudantes, académicos, líderes universitários, indústria e governos.

Os indicadores de desempenho do *ranking* estão agrupados em cinco áreas: Ensino (ambiente de aprendizagem); Pesquisa (volume, receita e reputação); Citações (influência da pesquisa); Perspectiva internacional (funcionários, alunos e pesquisa); e Renda da indústria (transferência de conhecimento)[7].

A lista das dez primeiras instituições classificadas pelo *website*[8], consultadas em 09/04/2021, encontram-se na Tabela 1.

Tabela 1 – Classificação de Universidades Portuguesas «Fonte: https://www.timeshighereducation.com»

Classificação	**Nome**	**Número Estudantes**	**Estudantes/*Staff***	**Estudantes Internacionais**	**Média Estudantes Femininos e Masculinos**
1001+	Universidade Trás-os-Montes e Alto Douro	6.515	11,4	4%	56:46
1001+	Instituto Politécnico do Porto	18.259	16,8	6%	48:52
801-1000	Universidade do Minho	18.504	18,2	13%	55:45
801-1000	Universidade do Algarve	8.008	12,9	19%	57:43
601-800	ISCTE-Instituto Técnico de Lisboa	8.868	23,3	12%	50:50
601:800	Universidade de Coimbra	21.332	16,6	18%	57:43
601-800	Universidade da Beira Interior	7.180	16,1	18%	52:48
601-800	Universidade de Aveiro	9.931	16,5	13%	51:49
501-600	Universidade de Lisboa	49.019	17,9	15%	52:48
401-500	Universidade do Porto	32.586	17,6	18%	55:45

[7] https://www.timeshighereducation.com/sites/default/files/breaking_news_files/the_2021_world_university_rankings_methodology_24082020final.pdf [Consult. em 23/08/2021].
[8] https://www.timeshighereducation.com/world-university-rankings/2021/world-ranking#!/page/0/length/25/locations/PT/sort_by/rank/sort_order/desc/cols/stats [Consult. em 23/08/2021].

Os dados em formato JSON[9] no endereço[10] do "Portal dos dados abertos da administração pública" contém as informações referentes aos cursos de cada instituição. Essas informações também foram consideradas no âmbito deste trabalho.

Para a pesquisa das informações de saídas profissionais, a limitação necessária é aplicação de filtros, limitando as buscas ao território português e de preferência, que o portal esteja inserido na cultura portuguesa. Neste quesito, existem vários *websites* de ofertas de emprego, tais como "portal emprego"[11], "jooble"[12], "alerta emprego"[13], "indeed"[14], "net empregos"[15], "sapo empregos"[16], "empregos online"[17], entre outros. Optou-se, no âmbito deste exercício académico, por utilizar o *website* "net-empregos.com", por ser esse o que mais ofertas de trabalho apresentou em pesquisas manuais neste ano de 2021, o que permite construir uma bateria de saídas profissionais mais completa. Além disto, o "net-empregos.com" é um *website* estritamente português. Para utilização futura, pode ser configurado outros *websites* de emprego, ampliando este trabalho para além dos limites deste exercício académico, trazendo maior alcance e escalabilidade da solução proposta.

1.5 Organização do documento

O documento deste trabalho está organizado como se segue:

No Capítulo 2, é apresentada a fundamentação teórica que está organizada em três partes. Na primeira parte, explora-se a opinião de diversos autores a respeito da carreira profissional, a importância e os fatores que influenciam a sua escolha, e ajuda a perceber as necessidades profissionais permitindo a seleção dos atributos mais relevantes que serão apresentados no portal. Na segunda parte, é explicada de forma resumida o que é a *internet* e como foi criada, e suas características de navegabilidade e utilização, com destaque para as suas limitações no que diz respeito à dificuldade de sistematização da informação, e respetiva tomada de decisão. Introduz também a temática da *web* semântica, como forma de simplificar essas tarefas e aliar a interpretação dos dados por máquinas, para além das pessoas. A última parte reforça conceptualmente que as ontologias são consideradas a base para *web* semântica

[9] https://pt.wikipedia.org/wiki/JSON [Consult. em 23/08/2021].
[10] https://dados.gov.pt/pt/datasets/r/59ed02b9-410c-4f68-81ef-a3755ca66400 [Consult. em 23/08/2021].
[11] xww.portalemprego.pt [Consult. em 23/08/2021].
[12] pt.jooble.org [Consult. em 23/08/2021].
[13] alertaemprego.pt [Consult. em 23/08/2021].
[14] pt.indeed.com [Consult. em 23/08/2021].
[15] net-empregos.pt [Consult. em 23/08/2021].
[16] emprego.sapo.pt [Consult. em 23/08/2021].
[17] empregosonline.pt [Consult. em 23/08/2021].

na criação de padrões que podem ser lidos também por máquinas. Esses padrões são definidos com sendo em formatos abertos e possuem recursos únicos e interação, representados por triplos (sujeito, predicado e objeto). Os conceitos de URI, RDF, RDFS, a linguagem OWL e SPARQL são explicados e exemplificados de modo a permitir a aprendizagem destes termos. Quanto as ontologias em si, são demonstradas tanto a forma de construí-las, quanto o mapeamento e as diversas formas de abordagem no instanciamento dos dados.

No Capítulo 3 são apresentado alguns exemplos de utilização das ferramentas da *web* semântica identificados na literatura e são descritas as considerações importantes de cada um para a construção deste trabalho. É apresentado um resumo das suas particularidades de forma a permitir uma vista abrangente da aplicabilidade da *web* semântica dentro e fora do contexto académico.

No Capítulo 4 é explicada qual a metodologia utilizada no trabalho de investigação, com destaque para a construção do artefacto – o portal, incluindo a criação e mapeamento das ontologias, a construção das tabelas relacionais, a extração dos dados na *web*, o instanciamento dos dados de acordo com as ontologias e a apresentação dos dados em formato aberto.

No Capítulo 7 são apresentadas as conclusões deste trabalho.

No Capítulo 8 são apresentadas as sugestões para trabalhos futuros.

No Capítulo 9 é apresentada a bibliografia utilizada na construção do presente trabalho, seguido pelos anexos.

2 Fundamentação Teórica

Neste capítulo serão apresentados os conceitos e fundamentação teórica necessários para a compreensão da proposta de dissertação.

Como retaguarda do trabalho, serão evidenciados os fatores que influenciam na decisão da escolha de carreira profissional. Os fatores intrínsecos serão mencionados mas não fazem parte do âmbito do trabalho, porém os fatores extrínsecos serão discutidos em detalhes.

Na seção seguinte será retratado o uso da *internet* no processo de pesquisa e análise de dados seguindo-se uma seção sobre a evolução natural da *internet* para a *web* de dados e como a *web* semântica poderá nesse caso concreto, auxiliar o aluno de forma inteligente e rápida, no processo de tomada de decisão da carreira profissional.

Na última seção deste capítulo, será detalhado de forma mais técnica a integração de dados bem como as tecnologias disponíveis e as que serão utilizadas no desenvolvimento deste trabalho.

2.1 Carreira Profissional

Diversos fatores de escolha de carreira foram evidenciados ao longo do tempo por diferentes autores. A escolha de carreira com foco em fatores ambientais ou de personalidade foi mencionada por [1], [2], [3] e [4]. Da mesma forma, alguns estudos sobre a escolha da carreira foram limitados apenas às escolas secundárias, ignorando o nível universitário [5] e [6]. Nesses estudos, comprovou-se que a seleção da carreira certa para o candidato certo pode ser alcançada quando a decisão foi tomada com base no conhecimento da avaliação de vários fatores influentes. Entretanto, não houve um procedimento claro sobre como os alunos podem adotar para a seleção de carreira, ao contrário dos estudos existentes de [7], [8] e [9], que enfatizaram o facto de que deve-se considerar a influência dos pares, família, personalidade e oportunidade sobre a escolha de carreira entre estudantes universitários.

2.1.1 Escolha da carreira profissional

O termo carreira originou-se das línguas latina e francesa. Segundo [10], significa ocupação, seja uma atividade social ou económica que pode ser aceite por alguém durante o processo de aprendizagem em instituição académica ou noutro lugar e prossegue ao longo da vida. Os autores em [11] definiram o termo carreira como uma série de cargos, deveres, tarefas e experiência profissional acumuladas por um indivíduo.

A decisão de estabelecer uma carreira refere-se a uma condição em que o indivíduo encontra dificuldades em selecionar uma determinada ocupação [12]. Observa-se que é uma decisão crítica feita ao selecionar a profissão ou ocupação que atenda às necessidades do indivíduo. A decisão adequada na tomada de decisão de carreira fornece uma estrutura inovadora que serve como solução eficaz para os desafios futuros. Afinal, escolher uma carreira é uma ação significativa e inevitável feita por alguém na vida. Escolher a opção errada e depois mudar de uma carreira para outra tem efeitos psicológicos negativos.

A escolha da carreira tem muitas implicações para a vida, podendo implicar na fonte de rendimento que gera satisfação, no reconhecimento da comunidade e no sucesso em geral. Não é uma tarefa simples e envolve um difícil processo de tomada de decisão. É uma questão universal por natureza e tem um impacto duradouro num indivíduo.

Recentemente, a escolha da carreira tem recebido atenção especial entre os investigadores ao examinar as suas variáveis influentes. Vários estudos foram realizados em todo o mundo e relataram a relação significativa entre as variáveis e sua influência na seleção de carreira. Nos resultados de um estudo apresentado em [13], envolvendo a escolha de carreira na identidade de carreira, consciência sobre a ocupação e desempenho académico revelaram que todas as variáveis são interdependentes e estes resultados foram considerados precursores importantes nessa área de estudo.

2.1.1.1 Influência de amigos na escolha da carreira

O autor em [14], explica que o grupo de amigos é um moderador importante na formação da percepção geral, comportamento e atitude do indivíduo por meio do processo de socialização que resulta na adoção de um determinado estilo de vida ou caráter que determina a decisão em vários assuntos. Considera-se grupo de amigos, as pessoas que fazem parte da convivência de determinado indivíduo e se enquadram na mesma faixa etária, comportamento, interesse, área geográfica e situação económica [14]. No estágio inicial do crescimento infantil, os pais desempenham um papel significativo na formação de comportamentos e atitudes mas mais tarde, num determinado estágio do desenvolvimento infantil, os amigos assumem o controlo para estimular a atitude e a decisão em vários assuntos, incluindo a carreira. Os amigos são uma fonte de pensamento crítico e ligação para a escolha de carreira e busca de emprego [15].

A influência dos amigos na escolha de carreira pode ser positiva ou negativa, dependendo da ligação e exposição do grupo de amigos. Se o aluno for aconselhado por pressão de colegas a fazer a escolha errada de carreira, os alunos encontrarão dificuldades em lidar com

o currículo, o que levará à insatisfação com a carreira escolhida. Assim, o *status* social do sucesso académico individual, a obtenção de empregos de boa reputação e ganhos elevados estão associados a uma estreita influência dos amigos. Os estudos existentes relataram que o grupo de amigos desempenha um grande papel significativo na modificação do comportamento da atitude e personalidade dos indivíduos em relação à decisão também sobre questões académicas [16] e [17].

2.1.1.2 Influência da família na escolha da carreira

O desenvolvimento do comportamento e da atitude dos jovens é controlado inicialmente a nível familiar, onde os jovens são inspirados por planos e estratégias futuras [18]. A influência da família pode ser positiva ou negativa dependendo da consciência e exposição à situação global.

O nível educacional dos pais, sua profissão e rendimentos são identificados como fator importante [19]. Os pais são uma fonte de aconselhamento no desenvolvimento de carreira. Por exemplo, se uma criança admira as habilidades dos pais no tratamento de pacientes em centros de saúde, espera-se que ela se interesse por estudos de ciências da saúde. Da mesma forma, se um dos pais for professor, a criança também pode ser motivada a pensar em escolher essa carreira. De acordo com [20], a família pode pressionar os seus filhos a escolherem a carreira que corresponda aos seus planos. Quando uma família em particular, a título de exemplo, está envolvida em grandes atividades comerciais e deseja que os seus filhos assumam o controlo da operação e da administração dos negócios, eles possivelmente irão preparar seus filhos para estudarem matérias relacionadas com gestão ou administração de empresas.

Às vezes, a decisão dos filhos sobre a escolha da carreira pode ser diferente do conhecimento adquirido com o estilo de vida dos seus pais. Se a natureza da ocupação dos pais os afasta de passar algum tempo e socializar com seus filhos em casa, os filhos podem decidir sobre a escolha de uma futura carreira que vai contra a ocupação da sua família [21].

Portanto, é importante que os pais estejam atentos às mudanças económicas e tecnológicas globais para que os conselhos aos seus filhos correspondam a estas mudanças. Da mesma forma, os resultados de um estudo apresentado em [22] relataram que a escolha de carreira individual é amplamente influenciada pelos pais que planeiam e moldam os seus filhos de acordo com sua determinação.

Os autores em [23] dizem que é mais provável a uma criança vinda de um ambiente onde recebe apoio dos pais e convive harmoniosamente, que a sua escolha profissional seja

ditada por eles. Nesse cenário, a aspiração ocupacional da criança tem mais probabilidade de ser influenciada pela profissão dos pais. Em sua análise, concluiu que os países desenvolvidos direcionam os seus alunos para carreiras de acordo com as necessidades do país.

2.1.1.3 Influência de oportunidades na escolha da carreira

A escolha da carreira profissional é muito influenciada pelas oportunidades associadas a uma carreira específica. Os alunos podem mostrar ambição e habilidade em seguir determinada carreira, porém, se não forem orientados a fazer as escolhas certas no momento certo, os seus sonhos acabarão por ser irrealistas [24]. Ter uma exposição para as oportunidades disponíveis faria com que o aluno tivesse uma boa chance de selecionar a melhor carreira que corresponda às suas competências e habilidades.

As oportunidades podem ser na forma de qualificação de acesos à vida académica, acompanhamento de empregos ou atividades no campo mais prático. Entretanto, a oportunidade potencial mais elevada será a oportunidade de emprego [25].

Prever as necessidades do mercado para certas carreiras é muito desafiador, uma vez que é determinado por vários fatores. A carreira, de um modo geral, pode ser monetizável por um determinado período, mas depois de alguma inovação tecnológica, ela está sujeita a perder essa capacidade de monetização. Entretanto, uma investigação crítica sobre as necessidades atuais dos empregadores pode ser feita por meio de uma comparação entre a oferta e a procura dos profissionais [26]. Isso pode ser observado rapidamente olhando anúncios de emprego de diferentes fontes. Esse esforço inicial pode ser útil para prever a procura de determinado campo de estudo por alguns anos.

Foi observado também que os alunos tomam uma decisão sobre a escolha da carreira, embora não tenham informações claras sobre o que existe no mercado global [27].

2.1.1.4 Influência da personalidade na escolha da carreira

A personalidade dos alunos está entre os fatores que influenciam a escolha da carreira. Os alunos escolhem caminhos de carreira que se encaixam exatamente com a sua personalidade. O termo personalidade refere-se ao processo psicológico proveniente do cérebro humano que indica o caráter individual [28].

Personalidade gera comportamento e atitude que se tornarão a identidade de alguém e é a maneira de distinguir um dos outros ao interagir com as pessoas. Mudanças de personalidades implicam na variação de decisão crítica em alguns assuntos valiosos que têm impacto na vida.

A autoconfiança dos alunos determina o seu plano futuro, escolhendo eles mesmos o caminho certo em direção aos seus objetivos futuros. Estudos indicam que os alunos com personalidade investigativa normalmente preferem cursos relacionados com ciências exatas, enquanto aqueles com personalidade artística geralmente decidem selecionar cursos focados nas ciências sociais [29].

Compreender as personalidades dos alunos e, em seguida, combiná-los com um tipo de carreira adequado pode melhorar a satisfação na carreira. Assim, a consciência da própria personalidade é vital para que o aluno faça a escolha certa para a sua carreira e é um construto fundamental para uma valiosa escolha de carreira que determina o sucesso futuro. Segundo a teoria apresentada em [30], a personalidade individual pode influenciar alguém a escolher uma determinada carreira.

Os indivíduos buscam um ambiente de trabalho que melhor se adapte aos seus interesses e capacidades. Em relação a essa teoria, as pessoas têm baixa ou alta identidade. Aqueles com baixa identidade normalmente selecionam carreiras incompatíveis e são caracterizados por mudarem de uma profissão para outra diferente daquela com alta identidade. A teoria explica ainda que o ser humano é acompanhado por vários traços de personalidade, especificamente:

- Realista,
- Investigativo,
- Artístico,
- Social,
- Empreendedor e
- Conversivo.

Estas características variam de acordo com o tipo de tarefa atribuída a indivíduos que dependem da habilidade e competência para realizar uma determinada atividade.

Os autores em [31] investigaram a relação entre personalidade e tomada de decisão de carreira, realçando a ligação entre personalidade efetiva e processo de tomada de decisão maduro. O termo personalidade efetiva é definido conceptualmente como um grupo de características únicas de indivíduos que orientam e conduzem à tomada de decisão crítica. Concluiu-se que quanto maior for a personalidade efetiva, maiores as possibilidades de possuir e usar assertividade, autoestima e confiança.

2.1.1.5 Outros fatores na escolha da carreira

Existem ainda outros fatores na escolha da carreira que tornam essa decisão num dos maiores dilemas e desafios para a vida de qualquer pessoa. Além dos fatores já mencionados, existem outros de menor peso mas que estão intrinsecamente interligados.

Por exemplo, perspetivas financeiras influenciam preferencialmente a escolha de carreira por homens e em como eles têm que atender a despesas da família, enquanto as mulheres mostram mais preocupação com os valores sociais e utilidade. A escolha da carreira é, por isto, determinada também pelo seu nível potencial de rendimento, natureza de funções e, consequentemente, pelo legado que é deixado nas funções a desempenhar.

O reflexo disso manifesta-se numa escala maior na prosperidade económica de uma nação. Indivíduos que são desajustados nos seus locais de trabalho tendem a ser menos produtivos e eficientes e, portanto, incapazes de atingir os seus objetivos. O conceito que define ocupação como um meio de vida, que tem o poder de mudar personalidades, determinar o social *status*, prever ganhos esperados ou determinar grupos sociais, entre outros. Alguns fatores como aptidão, circunstâncias de vida e desempenho académico também foram comprovados como determinantes na escolha de carreira [32]. Dado a sua complexidade, é então um ponto a ponderar sobre como as decisões de carreira são tomadas.

Cada aluno numa determinada conjuntura em sua vida tem que fazer uma escolha em relação à sua carreira. É obrigatório que os alunos façam uma escolha correta, afirmado em [33]. Isso vai fazê-los mais equilibrados e estáveis. Consequentemente, isso levará a um melhor resultado para a sociedade.

2.1.2 Considerações na escolha da Carreira Profissional

Infelizmente, as escolhas de carreira são feitas com pouca consciência do mundo real segundo os autores em [34] e [35]. Segundo eles, os alunos tomam decisões cruciais num estágio quando eles podem não estar totalmente informados de suas escolhas, ou então em circunstâncias inevitáveis que os impedem de perseguirem seus objetivos.

Estar interessado numa profissão em particular é muito importante na tomada de decisão. Se um aluno é forçado a uma carreira, ele pode exibir baixa autoestima e baixo desempenho. O autor em [36] relata que vários estudos indicaram uma relação positiva entre interesses e escolha de carreira.

As profissões podem ter vários graus de aceitabilidade em diferentes culturas o que também influencia a escolha de carreira de um indivíduo [37].

Segundo o autor em [33], o papel da escola é fornecer orientação e também incentivar os alunos a continuar com a educação e a não desistir.

Em muitos casos, as decisões que envolvem a escolha de cursos, cursos de especialização e as carreiras subsequentes são igualmente exaustivas e difíceis para alunos completando a escolaridade e prosseguindo para faculdade [38].

2.2 A *Internet* como ferramenta de pesquisa

A *Internet* é um mundo por si só, de conhecimento disperso e acessível por seres humanos, que podem recolher e consultar informações, analisá-las e fazer as suas próprias conclusões. Para a escolha da carreira isso não será diferente. Nesta seção será abordado o processo de pesquisa na *Internet*.

Segundo os autores em [39], "[...] o processo de pesquisa, nos moldes tradicionalmente praticados, por vezes é limitado em função de custo, tempo, dispersão geográfica ou intensidade de trabalho. Tais barreiras podem ser exponencialmente resolvidas com o uso da tecnologia *Internet*. Ela oferece um novo cenário tecnológico para a recolha e tratamento de dados necessários para a realização de pesquisas [40] e [41]. Considere-se também que a acessibilidade universal das tecnologias de informação significa que a população de utilizadores pode ser extremamente diversa e rica em termos de experiências, características, habilidades e retornos [42]."

Sendo considerada como uma das tecnologias de maior influência em difusão de informações e interatividade em [43] e [44], a *internet* posiciona-se como uma ferramenta importante para a aquisição de dados e apresentação de resultados, revolucionando a maneira como se obtém informações.

As propriedades da *Internet* permitem a realização de pesquisa tanto de um processo sequencial como também de um processo em paralelo, fornecendo acesso imediato a diferentes tipos e formato de informações.

Tais fatores podem permitir a evolução de novos caminhos no processo de pesquisa, afetando o comportamento do ator envolvido. Desta forma, o processo de pesquisa na *Internet* assemelha-se ao processo tradicional de pesquisa, envolvendo as fases de preparação, recolha

de informações, tratamento dos dados e divulgação dos resultados, porém estas etapas do processo misturam-se dinamicamente e tornam-se interativas e simultâneas. [39].

2.2.1 Dados e informações digitais

Segundo o autor em [45] "em um estudo realizado em 2003 por investigadores da Escola de Gestão de Informação e Sistemas da Universidade de Berkeley, estimou-se que a humanidade tenha acumulado 12 exabytes[18] (12 x 10^18) de dados até o momento anterior aos computadores virarem *commodities*, na década de 90. No entanto, este mesmo estudo mostrou que somente no ano de 2002, a humanidade produziu mais de 5 exabytes em equipamentos de armazenamento ótico, magnético, filme e impressão. Esta produção é equivalente à produção de 37 mil novas bibliotecas do tamanho da Biblioteca do Congresso Americano. Destes 5 exabytes produzidos, 92% foram armazenados em dispositivos magnéticos, a maioria em discos rígidos, o que aponta para uma incrível democratização da informação [46]".

Existe uma ligação entre a informação e o desenvolvimento da sociedade pós-moderna. Os membros do G7 – Canadá, França, Alemanha, Itália, Japão, Reino Unido e Estados Unidos podem ser qualificados como sociedades da informação, pois pelo menos 70% dos seus PIBs dependem de bens intangíveis (bens relacionados à informação) [46]. Desta forma, conclui-se que o funcionamento e o crescimento destas nações dependem da constante geração e consumo de dados de forma massiva.

Num estudo mais recente, foi destacado que de 2006 a 2010, o número de dados digitais gerados cresceu de 166 exabytes para 988 exabytes. Esta grande quantidade de dados que tem sido gerada já alcançou a casa dos zettabytes(1000 exabytes ou 10^24) em 2014 [46].

Como descrito anteriormente, mais de 90% destes dados estão armazenados em discos rígidos, onde centenas de milhões de computadores, ininterruptamente, processam estes dados em busca de informação útil e relevante e, às vezes, nova. Estima-se que nesse ano de 2020, o volume de dados chegará a 40 zettabytes.

Este incrível crescimento de dados faz com que seja fundamental uma cuidadosa análise destes dados e a compreensão sobre quais os tipos de dados que têm sido gerados nesta sociedade da informação.

[18] https://pt.wikipedia.org/wiki/Exabyte [Consult. em 23/04/2021].

Os dados digitais estão, de alguma forma, estruturados e têm sido gerados com o intuito de prover informações úteis e gerar novos conhecimentos.

É importante frisar que muitos desses dados podem ser descobertos e acedidos, tanto por seres humanos quanto por máquinas, através da *Internet*.

Em 1989, o físico inglês, Sir Timothy John Berners-Lee, no CERN do francês *Conseil Européen pour la Recherche Nucléaire* (ou Centro Europeu de Pesquisas Nucleares), inventou a WWW (*world wide web*), através da proposição de três tecnologias fundamentais:

1 HTML (*Hypertext Markup Language*);
2 Servidor HTTP (*Hypertext Transfer Protocol*);
3 URI (*Unified Resource Identifier*).

Tais invenções foram motivadas pela necessidade de facilitar a partilha de documentos entre os investigadores e visitantes do CERN. Com isso, tanto o lançamento do *browser* "X *Windows Mosaic* 1.0" quanto a primeira conferência sobre a WWW estimularam a desenvolvimento da *web*.

Em maio de 1994, houve a primeira conferência internacional sobre a WWW (*First International Conference on World Wide Web*), em Genebra. Esta conferência simboliza a grande popularização da *web*, onde nela foi anunciado o consórcio que cuida dos padrões e tecnologias relacionadas ao desenvolvimento da *web*, o W3C (*World Wide Web Consortium*).

Ainda nesta conferência, Tim Berners-Lee proferiu uma palestra sobre a necessidade de semântica na *web*. Segundo Tim Berners-Lee, a forma que os documentos estavam estruturados (através de nós e *links*) fazia com que apenas serem humanos pudessem entender o significado contido nelas. Tal impossibilidade fazia com que máquinas não pudessem aceder e obter significado dos documentos.

De forma mais detalhada, os documentos utilizam a linguagem HTML, que é uma linguagem que faz apresentação de hipertextos. Os hipertextos são documentos que possuem *links* e nós (ou pontos de ligação) com outros documentos, permitindo assim a navegação entre os mesmos. No entanto, os *links* que existem para relacionar os documentos não tinham nenhum tipo de característica que os diferenciassem um do outro, de tal forma que não é possível para as máquinas distinguirem o significado entre uma relação e outra.

É importante frisar que tais problemas ficam mais evidentes nos dias de hoje, pois se calculam dezenas de milhares de milhões de páginas *web* disponíveis e mais de 1 zettabyte de

dados. Esta quantidade de documentos torna praticamente impossível o acesso e a pesquisa por informação de forma eficiente e consistente pelos seres humanos, fazendo com que haja a necessidade de *softwares* recolherem informações e processarem atividades para os humanos [47].

Com isso, a semântica acessível por máquina é potencializada através da especificação de documentos *web* numa linguagem que permita que os *links* sejam criados com valor ao seu relacionamento. Isto faz com que os recursos possuam uma semântica associada, permitindo a execução automática de atividades como compra de produtos personalizados, negociação por pacotes turísticos, agendamento de consultas, entre outras.

Em 1999, Tim Berners-Lee publicou um livro que se intitulava: "*Weaving the Web*", onde o termo *web* semântica apareceu pela primeira vez. Finalmente, em 2001, um artigo publicado na revista Scientific American marcou o início da pesquisa relacionada à *web* semântica [48]. Neste artigo, Tim Berners Lee aborda características da *web* semântica, propondo as camadas da *web* semântica e descrevendo como poderiam funcionar.

De forma mais detalhada, a *web* semântica visa a utilização de recursos provenientes da Inteligência Artificial (como agentes inteligentes e representação de conhecimento), Engenharia de *Software* (como *frameworks* e plataformas), Computação Distribuída (como *web services*), entre outras, para executar atividades na *web* que antes só eram possíveis por agentes humanos [48].

A *web* semântica estende a *web* clássica, provendo uma estrutura semântica para páginas *web*, a qual permite que tanto agentes humanos quanto agentes de *software* possam entender o conteúdo presente em páginas *web*. Dessa forma, a *web* semântica permite um ambiente onde agentes de *software* podem navegar através de páginas *web* e executar tarefas sofisticadas. Em outras palavras, a *web* semântica é necessária para expressar informações de forma precisa, podendo tais informações serem interpretadas por máquinas e dessa forma permitirem que agentes de software possam processar, compartilhar, reusar, além de poder entender os termos que estão sendo descritos pelos dados [49] e [50].

2.3 A *web* semântica

Como dito pelos autores em [51], "A informação na *web* é tipicamente representada em linguagem natural permitindo que ela seja compreendida por pessoas. Contudo, para prover informação de forma que computadores também possam compreendê-la (e extrair o seu significado) é necessário representá-la de forma sistemática e semântica. A *web* semântica foi

o nome utilizado para introduzir a nova geração de tecnologias que tem como objetivo representar a informação de uma maneira na qual computadores sejam capazes de interpretá-la. Além disso, através desta representação, as pesquisas em *web* semântica propõe tecnologias para automação, integração e re-uso da informação mesmo considerando diferentes plataformas de desenvolvimento, sistemas operativos, protocolos de rede, e outras variações de tecnologia [49]. Atualmente este é um dos principais tópicos de investigação das comunidades de Inteligência Artificial e de *Internet* [53].

Segundo o autor em [54], as ontologias são consideradas a base da *web* semântica oferecendo uma linguagem expressiva e formal para gerar informação que pode ser interpretada por computadores. Estas ontologias podem ser combinadas, compartilhadas, modificadas e utilizadas para anotar "semanticamente" diferentes tipos de recursos como, páginas *web*, documentos, unidades de armazenamento (digitais ou não), além de outros recursos [55]. Dessa forma, as ontologias oferecem a possibilidade de incluir significado na informação descrita na *web* permitindo que os computadores "raciocinem" (realizem inferências) em cima dos dados disponíveis na *web* de forma mais "inteligente". Segundo o autor em [56], o uso de ontologias e o desenvolvimento de serviços *web* para processar a informação disponível na *Internet* está transformando a *web* da informação (*web* tradicional) na *web* do conhecimento. Como mostra a Figura 1, a visão da *web* semântica é criar a *web* do conhecimento onde a informação está distribuída em diferentes repositórios e anotada utilizando ontologias interconectadas.

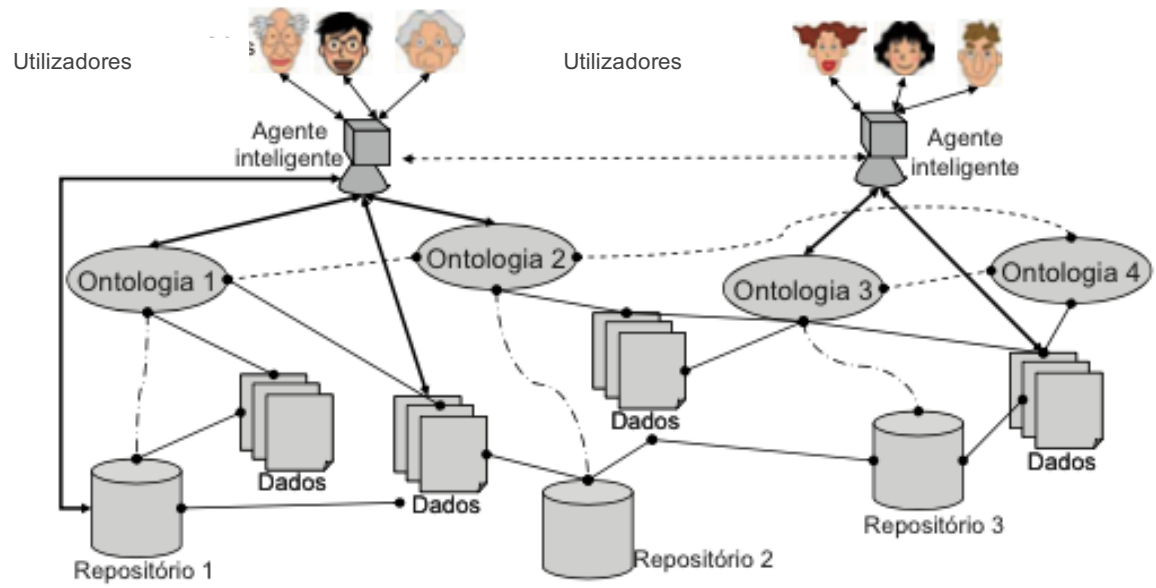

Figura 1: A Visão da web semântica «Fonte: autor em [56]»

Através destas ontologias, agentes inteligentes podem aceder, compartilhar e trocar informações de maneira eficiente facilitando o desenvolvimento de serviços que agregam dados de diferentes localidades. Essa mudança de paradigma na *web* permite que computadores e pessoas trabalhem cooperativamente de maneira muito mais eficiente [57].

O autor em [52] ressalta ainda que a *web* semântica propõe tecnologias que dão o "poder" aos computadores através da criação de padrões, protocolos e linguagems formais que facilitam o processamento da informação na *web* de forma automática e inteligente. Contudo, para o utilizador final (internauta), a *web s*emântica ainda é muito complexa para ser utilizada. Isso ocorre porque a *web* semântica foi concebida para facilitar o consumo da informação pelos computadores. Portanto, as aplicações normalmente não dispõem de uma interface amigável que proporcionem maior liberdade ao utilizador final para criar e compartilhar a informação.

2.3.1 Integração de dados

Como dito anteriormente, a invenção da *internet* e a revolução dos computadores levou o acesso ao dados digitais a bilhões de pessoas. Assim, manipulá-los e obter o máximo do seu potencial é algo desejado por muitos. No entanto, diferente do isolamento dos sistemas de informação tradicionais, atualmente tem crescido a necessidade por sistemas integrados e interoperáveis.

Segundo o autor em [61], "a integração de dados promove a manipulação transparente dos mesmos entre as múltiplas fontes [62]. Entretanto, o processo de integração dos dados é cheio de desafios: por questões de sistemas, como montar uma estrutura de comunicação entre as bases de dados; por razões lógicas, como a heterogeneidade semântica das múltiplas bases de dados; ou pelo acesso, organização e gestão dos dados [63].

As fontes de dados podem ser heterogéneas na sintaxe, no esquema ou na semântica. A heterogeneidade sintática é resultado de diferentes modelos e linguagens de dados. Já a heterogeneidade esquemática é fruto de diferenças estruturais, enquanto que a semântica é causada pela diferença do significado ou interpretação do dado num determinado contexto [62].

Segundo o autor em [64], integrar dados oriundos de diferentes fontes é um processo que consiste de três tarefas. De seguida essas tarefas serão apresentadas e discutidas:

1) Alinhamento de esquema: Essa tarefa consiste em identificar as diversas fontes de dados, bem como seus atributos, estruturas, vocabulários e completude. Aqui também é importante notar como se dará o acesso aos dados e se há alguma restrição em termos

de privacidade. O componente principal do alinhamento de esquema é o Mapeamento de esquema.

2) Mapeamento de esquema: É no mapeamento do esquema onde ocorrerá a especificação de quais dados existem e de como os termos utilizados nas fontes de dados se relacionarão com o esquema que mediará a integração. Existem várias classes de mapeamento de esquema, entre elas as mais conhecidas, de acordo com [63], são:
   a) *Global-as-View* (GAV): Aqui, o esquema mediador, ou global, é definido como um conjunto de vistas sobre as fontes de dados. A principal vantagem do GAV é a sua simplicidade metodológica. Para reformular uma consulta basta refazer as ligações do esquema geral com as fontes de dados. A sua principal limitação está na adição e remoção de fontes de dados, que resultará na remodelação de todas as ligações;
   b) *Local-as-View* (LAV): Essa abordagem segue exatamente o oposto da anterior. Aqui, as fontes de dados servem de vistas para o esquema global, descrevendo-as da maneira mais precisa e independente, tanto quanto possível. Assim, a sua maior vantagem é a possibilidade de combinar os dados de diversas fontes, sendo fácil adicioná-los e removê-los. Entretanto essa facilidade cria uma maior complexidade na formulação das consultas;
   c) *Global-and-Local-as-View* (GLAV): O formalismo das duas abordagens anteriores é expresso na abordagem GLAV. Aqui, são realizadas duas consultas, uma a partir do esquema global e outra a partir das fontes locais, o que permite aproveitar os benefícios de ambas as abordagens. Muitos artigos sobre integração semântica de dados (que será vista nos tópicos posteriores) apresentam essas abordagens para guiar as metodologias de integração. Entretanto, com o passar dos anos e com a especialização das técnicas semânticas, a contextualização dessa abordagem tem caído em desuso.
3) Ligação dos registos: A principal tarefa da ligação dos registos é identificar e relacionar os mesmos indivíduos do mundo real nos diferentes registos das fontes de dados. Um caso especial de ligação dos registos é a deteção de duplicidades (presença do mesmo indivíduo numa única fonte de dados). Há diferentes métodos para a realização desses relacionamentos, entretanto eles geralmente são divididos entre:
   a) Ligação determinística: Na ligação determinística os registos idênticos de acordo uma ou mais variáveis são considerados como pertencentes ao mesmo indivíduo;

b) Ligação probabilística: Na ligação probabilística são executados algoritmos de proximidade entre os registos. Os registos que atingirem um nível aceitável de similaridade são considerados pertencentes ao mesmo indivíduo.
c) Fusão dos dados: A terceira tarefa consiste em unificar os registos que foram classificados como sendo o mesmo indivíduo. Quando esse processo é aplicado numa única base de dados ele é chamado de deduplicação.

2.3.2 Integração Semântica dos dados

O autor em [61] diz que "[...] de menor modo a integração semântica dos dados, na forma como geralmente é realizada, trate sobre o conteúdo semântico, é sabido que um maior esforço é necessário para a explicação e compreensão do contexto por trás do dados. Afinal é sabido que, para que os dados recolhidos pelos sistemas sejam plenamente úteis, é essencial que haja mais do que o acesso aos dados, é fundamental que exista o entendimento do seu significado, das suas relações e do contexto, portanto que exista um componente semântico. Essa necessidade é tornada mais evidente quando é preciso integrar ou ter uma visão geral de sistemas cujos modelos de dados são diversos, as propriedades heterógeneas e a terminologia vasta. Entretanto, comumente, durante o processo de modelação de base de dados relacionais e de folha de cálculo, o componente semântico fica implícito nos nomes das tabelas e das variáveis, dificultando o entendimento para qualquer um que não seja o criador do conjunto de dados [58]."

Existe então, a necessidade de ter sistemas que sejam capazes de expressar as ligações semânticas nos dados e entre os diferentes tipos de dados. Estes sistemas deveriam apresentar as nuances de contexto e, ao mesmo tempo, ser formal o suficiente para permitir que tanto *softwares* quanto seres humanos realizassem inferências baseadas no que está representado [58]. Quando, num processo de integração, há uma grande preocupação com os conceitos e as relações presentes nos conjuntos de dados, diz-se que há um processo de integração semântica. Nele, os dados semanticamente heterógeneos são integrados com uma menor perda de informação e uma maior recuperação do contexto [62].

No entanto, estabelecer ligações semânticas entre bases de dados pode ser uma tarefa difícil, principalmente se as fontes de dados forem grandes e complexas. Os autores em [58] apontaram como desafios no processo de integração semântica dos dados:

- O acesso aos dados;
- O acesso ao contexto e a interpretação real dos dados;

- O estabelecimento de relações significativas entre os dados.

Segundo os autores em [61], "[...] o acesso ao contexto e o estabelecimento das relações é algo que ainda dificulta o processo de integração, principalmente quando é necessário realizar um trabalho em larga escala. Ainda que nem todos os dados sejam disponibilizados na *web* e que a integração não seja feita com esse foco, o conceito de *web* semântica mudou a forma como se interage com o conhecimento representado, dando um novo fôlego para a área de integração semântica. A ideia de ter uma *web* conectada e com significado foi proposta por Tim Berners-Lee [57] e tem gerado contribuições até aos dias de hoje."

Com a *web* conectada, é possível ter a representação do mesmo objeto de diversas maneiras e ainda assim garantir a sua identidade e as relações que esse objeto tem com outros objetos. As tecnologias e padrões relacionados à *web* semântica permitem a criação de um ambiente onde as aplicações são capazes de fazer consultas e inferências sobre os objetos e suas relações. Ganha-se então a possibilidade de compreensão do conceito tanto por máquinas quanto por humanos.

Normalmente a *web* é compreensível por humanos, que conseguem interpretar o seu conteúdo por meio dos textos escritos nas páginas e do conhecimento prévio no contexto em que o texto está inserido, no entanto é incompreensível por máquinas devido a ausência de marcadores de informação e de conectores entre os conceitos presentes nos diferentes *websites*.

Para que isso ocorra, entretanto, é necessário que os dados e as relações entre dados sejam disponibilizadas num padrão. O nome desta coleção de fontes de dados inter-relacionados padronizadas presentes na *web* chama-se *Linked Data* [57].

2.3.3 Linked Data

Linked Data (LD) ou Dados Conectados é o nome dado à coleção de dados integrados presentes na *web*. Um exemplo de uma grande quantidade de dados conectados no formato preconizado pela *web* semântica é a DBpedia[19]. Nela, o conteúdo da Wikipedia[20] está disposto em *Resource Description Framework* (RDF)[21], um dos padrões da *web s*emântica, o que permitiu a integração com outras fontes de dados da *web* [57].

[19] https://www.dbpedia.org/ [Consult. em 23/08/2021].
[20] https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal [Consult. em 23/08/2021].\
[21] https://pt.wikipedia.org/wiki/Resource_Description_Framework [Consult. em 23/08/2021].

O formato em que esses dados estão apresentados também permite a execução de consultas, inclusive relacionadas ao conteúdo presentes em mais de uma fonte. Como resultado é obtida uma visão agregada sobre os conceitos modelados. Esta visão é tanto interpretável e reutilizável por máquinas, quanto enriquecida semanticamente para uma melhor compreensão por seres humanos, uma vez que a classificação dos dados se dá a partir de conceitos previamente definidos.

Um exemplo do uso dos conceitos de LD é o projeto *Linked Open Data* (LOD)[22], fundado em 2007 e que tem o suporte do grupo *Semantic Web Education and Outreach* (SWEO).

Há uma série de características que tornam os dados melhor ou pior classificados como LOD. Para isso, criou-se uma classificação de cinco estrelas. Quanto mais estrelas, mais fácil é o acesso e a possibilidade de uso dos dados por outras pessoas. São elas:

1) Disponibilização dos dados na *web* (independente do formato), mas numa licença aberta;
2) Disponibilização dos dados em formato compreensível por máquinas (como ficheiros de tabelas ao invés de imagens de tabelas);
3) Disponibilização dos dados em formatos não proprietários como o csv e não xls ou dbf, além de seguir os itens anteriores;
4) Utilização de formatos abertos apoiados pelo W3C para identificar objetos, além de seguir os itens anteriores;
5) Ligação entre os dados disponibilizados aos de outras bases públicas, além de seguir os itens anteriores.

Como pode ser visto na classificação apresentada, a construção de páginas *web* que seguem o princípio da *web* semântica exige o uso de diversos padrões, os formatos abertos apoiados pelo W3C, tais como RDF, URI, OWL e SPARQL [58].

2.3.4 Representação dos dados na *web* semântica

Segundo o autor em [45], o modelo RDF foi projetado para descrever recursos na *web*. Entretanto, é importante lembrar que a *web* é um espaço de informação no qual os itens de interesse precisam ser identificados. Assim, cada recurso possui um identificador único e global

[22] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData [Consult. em 23/08/2021].

para que o mesmo possa ser identificado na WWW, chamado de *Uniform Resource Identifiers* (URI), que é um padrão da *web* utilizado para identificar recursos abstratos ou físicos, sendo composto por uma sequência compacta de caracteres.

Para compreender isso é preciso entender a arquitetura da *web* que é composta por três bases fundamentais, como descritas a seguir [65]:

Recurso único: como explicado anteriormente, os identificadores são utilizados para identificar os recursos. O identificador é o URI, que permite uma maneira simples e única de identificar recursos na *web*, é caracterizado por três aspetos:

1) Uniformidade: utilização de recursos tanto no mesmo contexto quanto em contextos diferenciados;
2) Recurso: qualquer coisa que pode ser identificado por um URI, como um vídeo, imagem, serviço, documento, entre outros; e
3) Identificador: informação requerida para identificar e diferenciar um determinado recurso de qualquer outro.

Além disso, um URI pode ser classificado como:

- *Uniform Resource Locator* (URL), onde basicamente define um localizador/endereço para um determinado recurso através de um protocolo existente e
- *Unified Resource Name* (URN) representa um nome para um determinado recurso, garantindo unicidade e persistência de forma global mesmo quando o recurso não está disponível.

Finalmente, destacamos também o *International Resource Identifier* (IRI) que é uma generalização do URI. Diferente do URI, que é baseado os caracteres ASCII do inglês, *American Standard Code for Information Interchange*, o IRI amplia o número de caracteres Chineses (*kanji*) e Japoneses (*hiragana* e *katakana*), bem como os caracteres Cirílicos e Coreanos para que possam ser utilizados.

Interação: a *web* possui uma arquitetura cliente-servidor. A comunicação na *web* acontece através de protocolos padrões que permitem a troca de mensagens entre um servidor *web* que implementa este protocolo e um *browser* cliente que envia a solicitação para o servidor. O protocolo por defeito utilizado na *web* é o HTTP (*Hypertext Transfer Protocol*)[23], e como o próprio acrónimo diz, é um protocolo de transferência de documentos hipertextos (e.x. HTML).

[23] https://pt.wikipedia.org/wiki/Hypertext_Transfer_Protocol [Consult. em 23/08/2021].

Formatos: na comunicação cliente-servidor, o servidor irá retornar para o cliente (navegador) uma representação num determinado formato. Cada formato de representação retornado para o cliente conterá informação de metadados e dados. Os metadados são os atributos de cabeçalhos utilizados para identificar o formato de representação.

2.3.5 Triplos RDF

Os dados na *web* semântica são modelados e representados no padrão RDF, como discutido em [58]. Este modelo de dados permite uma descrição flexível dos objetos do mundo e ainda assim garante a descrição do contexto e a legibilidade por parte das máquinas (computadores) [59]. O RDF tem uma estrutura de triplos (composição de três elementos) baseada em grafos dirigidos, do tipo sujeito-predicado-objeto.

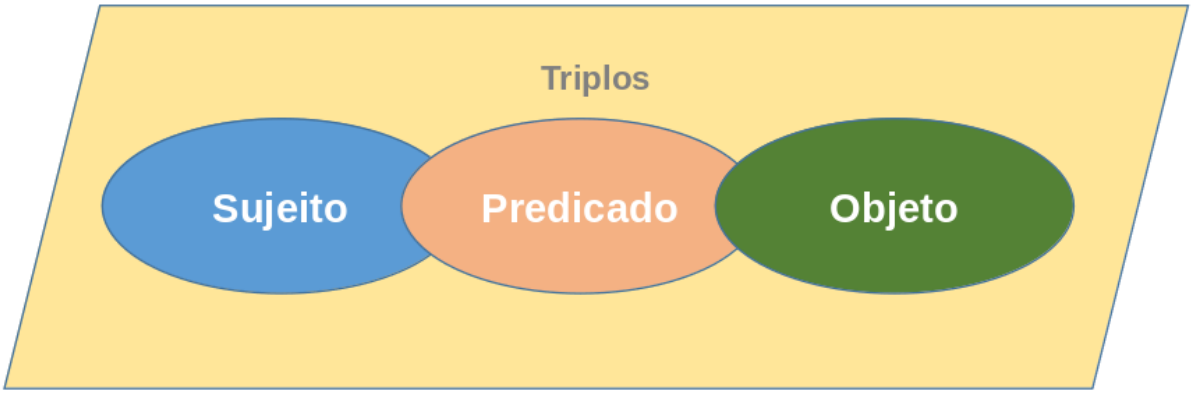

Figura 2: Grafo RDF (Triplos) «Fonte: elaboração própria»

Tanto o sujeito quanto o objeto do triplo identificam um componente e um conjunto de caracteres. Já o predicado do triplo especifica como o sujeito e o predicado estão relacionados.

Esta representação permite que um computador possa interpretar os dados representados por estes triplos. Contudo, para que isso seja completamente realizado precisamos garantir que cada um dos elementos do grafo sejam representados e referenciados de maneira única. E isso pode ser realizado utilizando os URIs [59].

Os literais (usados em conjunto com os tipos de dados) podem ser compreendidos como todos os valores identificados num grafo RDF que não possuem um URI associado. Os literais só podem ser aplicados a objetos, ou seja, nunca podem descrever sujeitos ou predicados.

Outro conceito importante do RDF é o conceito de Múltiplos Grafos. Ao se criar um documento RDF, pode-se adicionar outros grafos que se ligam ao grafo original, proporcionando assim catálogos de dados (do Inglês *Datasets*) conectados. O mais importante é que os múltiplos grafos podem ser acedidos por um único URI.

2.3.6 RDF *Schema*

A fim de adicionar descrição semântica ao modo de representação dos dados surgiu um complemento do RDF, o RDF *Schema* (RDFS). No RDFS há a adição de domínio e âmbito (*range*), além de propriedades, características e relações hierárquicas aos conceitos.

Segundo o autor em [45], o RDFS é um vocabulário para modelação de dados que amplia a expressividade do RDF para permitir mecanismos de descrição de taxonomias entre recursos e suas propriedades. Ou seja, o RDFS permite descrever grupos de recursos (também conhecidos como classes) e suas relações utilizando o conceito de triplos apresentado anteriormente.

Da mesma forma que o XML *Schema* é baseado no XML, o mesmo ocorre com o RDF *Schema*, que é baseado no RDF. Isto implica que o RDFS possui um URI para cada recurso e que também possui uma estrutura de triplos "sujeito, predicado, objeto".

É necessário destacar dois conceitos básicos presentes em RDFS:

1. Classe: este conceito é utilizado para descrever recursos num documento RDF. Toda classe RDF possui um URI associado e podemos definir propriedades para as classes RDF.
2. Propriedade: o papel de uma propriedade é definir relações entre sujeitos e objetos.

No entanto, mesmo o RDFS possui limitações na estrutura de representação semântica, segundo o autor em [60], uma vez que não apresenta construtores lógicos, como o encadeamento de propriedades. Não há possibilidade, por exemplo, de especificar restrições de cardinalidade. Assim, é impossível expressar que um ser humano possui exatamente dois progenitores. Assim, a fim de entender o RDF *Schema*, o W3C recomendou o uso da linguagem *Web Ontology Language* (OWL) [60].

2.3.7 A Linguagem OWL

A *Web Ontology Language* (OWL) é uma linguagem para definir e instanciar ontologias na WWW. Uma ontologia OWL pode incluir descrições de classes e suas respetivas propriedades e seus relacionamentos. OWL foi projetada para o uso por aplicações que precisam processar o conteúdo da informação ao invés de apenas apresentá-la aos humanos. Ela facilita mais a possibilidade de interpretação por máquinas do conteúdo da *web* do que XML, RDF e RDFS (RDF *Schema*), por fornecer vocabulário adicional com uma semântica

formal. A OWL foi baseada nas linguagens OIL[24] e DAML+OIL[25], e é hoje uma recomendação da W3C que mantém a sintaxe do RDF mas vai além da semântica proposta pelo RDF *Schema*, uma vez que se baseia numa lógica de descrição. Esse componente lógico permite a realização de interceções de conceitos, bem como a determinação da cardinalidade, como mencionado anteriormente.

Todo o projeto da OWL tem em vista o processamento de conteúdo semântico da informação na *web* além do aumento da facilidade de expressar o significado em formatos XML, RDF e RDFS. Por causa disto, pode ser considerada uma evolução destas linguagens em termos de sua habilidade em representar conteúdo semântico da *web* interpretável por máquinas. Como a OWL é baseada em XML, a informação pode ser facilmente trocada entre diferentes tipos de computadores, sistemas operativos e linguagens de programação e é por isso que pode criar padrões que forneçam um *framework* para uso intensivo por ferramentas na *web.*

Para a execução de consultas nos arquivos modelados em RDF (e consequentemente RDFS e OWL) foi desenvolvido o *SPARQL Protocol and RDF Query Language (SPARQL)*[26]. Essa linguagem compreende a sintaxe e o modelo de dados do RDF sendo também baseado num modelo de grafos [60].

SPARQL é uma recomendação do W3C a partir de Janeiro de 2008. Seu propósito é permitir que ficheiros RDF sejam consultados através de uma linguagem parecida com SQL. Permite ao utilizador combinar dados de ficheiros RDF provenientes de diferentes fontes. É uma linguagem orientada a dados, ou seja, recupera dados armazenados em ficheiros RDF.

A consulta é composta por:

1. Declaração dos prefixos, que são pseudónimos, abreviações das URIs. Geralmente os prefixos são utilizados para o acesso aos dicionários de tipos de dados e à próprias ontologias;
2. Definição do conjunto de dados, usando a cláusula *FROM*. No entanto o uso desta cláusula é opcional na escrita. Caso não seja especificado é assumido que a busca é realizada na ontologia do sistema da busca [60];
3. Cláusula de resultado, que identifica quais informações deverão ser retornadas a partir da consulta;

[24] https://pt.wikipedia.org/wiki/Ontology_Inference_Layer [Consult. em 23/08/2021].
[25] https://pt.wikipedia.org/wiki/DAML%2BOIL [Consult. em 23/08/2021].
[26] https://pt.wikipedia.org/wiki/SPARQL [Consult. em 23/08/2021].

4. Padrão da consulta, que especifica o que será consultado dentro do conjunto de dados;
5. Modificadores da consulta ordenam e reorganizam os resultados da consulta.

Na Figura 3 pode ser verificada a estrutura padrão de uma consulta SPARQL.

```
   # declarações de prefixo
PREFIX foo: <http://example.com/resources/>
   ...
   # definição de conjunto de dados
FROM ...
   # cláusula de resultado
SELECT ...
   # padrão de consulta
WHERE {
   ...
}
   # modificadores de consulta
ORDER BY
```

Figura 3: Estrutura padrão de uma consulta SPARQL «Fonte: autor em [60]»

Na Figura 4 são demonstradas algumas cláusulas específicas da SPARQL.

BASE: define a URI base de um recurso;

FILTER: aplica um filtro sobre as linhas recuperadas pela consulta;

LIMIT: limita a quantidade de linhas recuperadas da consulta;

OFFSET: permite que seja aplicado um deslocamento sobre o conjunto de linhas recuperadas pela consulta;

OPTIONAL: permite que uma linha seja recuperada mesmo que não exista o valor de uma propriedade do RDF;

PREFIX: cria um "apelido" para a URI de um ficheiro RDF/OWL.

Figura 4: Algumas cláusulas específicas da SPARQL «Idem»

Na SPARQL as variáveis são identificadas com os símbolos "?" e/ou "$".

No exemplo de uma consulta SPARQL, que pode ser vista na Figura 5 abaixo, pode-se distinguir claramente os componentes citados anteriormente.

```
SELECT ?x ?desc
WHERE {
   ?x a dbo:Band .
   ?x foaf:name ?name .
   ?x dbo:abstract ?desc .
   FILTER (lcase(str(?name)) = "iron maiden")
   FILTER (langMatches(lang(?desc), "PT"))
}
```

Figura 5: Exemplo de uma consulta SPARQL «Idem»

Ainda no seu trabalho, o autor em [61] explica que esses padrões são baseados em boas práticas. "[...] Ao publicar o seu artigo sobre LD, Tim Berners-Lee estabeleceu quatro regras ou princípios para o conteúdo na *web* semântica [66]. A seguir são apresentados os princípios com explicações baseadas no trabalho de [59]:

1) Utilização de URIs para nomear objetos: Este princípio evoca a utilização de URIs não apenas para documentos presentes na *web* ou nos meios digitais, mas também para objetos reais e conceitos abstratos;
2) Utilização de URIs HTTP, assim as pessoas poderão consultar os objetos: URIs HTTP são nomes, não endereços. Assim, este princípio leva a identificar objetos e extrair conceitos a partir desse padrão. Ao nomear os objetos, o seu acesso torna-se mais amplo;
3) Fornecimento de informação útil, utilizando os padrões (RDF, SPARQL), para pessoas que consultam uma URI: Ao consultar uma URI é esperado que as informações estejam em formatos voltados para a compreensão do contexto, como o RDF e a linguarem de consultas SPARQL e não num ficheiro compactado;
4) Inclusão de *links* para outras URIs, assim será possível descobrir novos objetos: Aqui é incentivada a prática de criação de relações com outros conceitos."

2.3.8 Ontologias: Definições, modelação e usos

Segundo o autor em [61], "ao representar os dados presentes numa rede de conhecimento, como por exemplo na *web*, por meio de conceitos e ligações entre os conceitos (e assim, com semântica), estes são apresentados em ontologias."

A ciência da computação considera a definição de ontologia com sendo uma "[...] especificação formal e explícita de uma conceptualização compartilhada" [67]. Onde, de acordo com [68], inclui características de:

1) "**especificação formal**": significa a possibilidade de compreensão por um computador, uma vez que é representada com um formalismo lógico-matemático;
2) "**explícita**": o conjunto de conceitos, relações e restrições são explicitamente definidos;
3) "**conceptualização**": é a representação num modelo abstrato de um domínio de conhecimento que identifica os conceitos relevantes do domínio; e
4) "**partilhada**": significa que o conhecimento modelado é aceito por um grupo.

Hoje é aceite também que as ontologias podem ser utilizadas para descrever semanticamente as informações a fim de tornarem o conhecimento explícito e auxiliarem no processo de integração das fontes de dados [69]. As fontes de dados, ao terem as suas correspondentes em ontologias, podem ser usadas na identificação e associação de correspondências semânticas dos conceito, de modo que o conhecimento explicitado nas ontologias assumirá uma forma de organização diferente ao apresentado nas bases de dados.

Segundo o autor em [50], uma ontologia é constituída pelo seguinte:

1. Um conjunto de conceitos essenciais resultantes da articulação do conhecimento básico presente num determinado domínio. Esses conceitos podem ser representados usando um vocabulário especializado.
2. O corpo de conhecimento, que descreve o domínio utilizando os conceitos essenciais. Ele é composto por:
3. Uma hierarquia (classe/subclasse) resultante das relações "como" entre conceitos.
4. Um conjunto de relações importantes entre conceitos além das relações "como" (por exemplo, "parte de").
5. Uma axiomatização de restrições semânticas entre esses conceitos e relações.

Pode-se definir então uma ontologia como um relacionamento de quatro elementos, representado por O = {C, R, I, A}, onde [70]:

- C: é o conjunto de classes que representam os conceitos num dado domínio de interesse;
- R: é o conjunto de relações ou associações entre os conceitos do domínio;
- I: é o conjunto de instâncias derivadas das classes, ou ainda, os exemplos de conceitos representados numa ontologia;
- A: é o conjunto de axiomas do domínio, que servem para modelar restrições e regras inerentes às instâncias.

Esses elementos que constituem uma ontologia são fundamentais para a criação de uma estrutura que represente o conhecimento de um domínio. A Figura 3 apresenta um exemplo de uma ontologia criada pelos autores em [58].

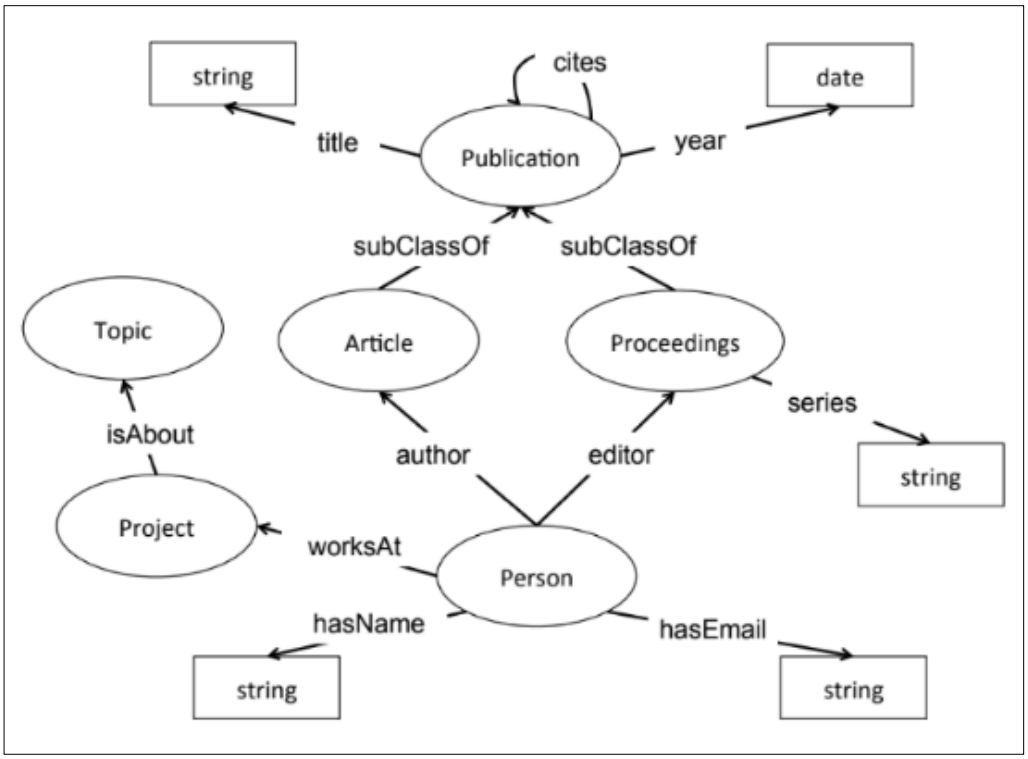

Figura 6: Representação de uma ontologia Semantic Web for Resource Community «Fonte: autores em [58]»

Na Figura 3, as classes são apresentadas por elipses. Elas são conceitos chaves de um domínio, podendo haver hierarquia como entre *Publication* e *Article.*

As propriedades de objeto são apresentadas como o rótulo das setas entre as classes, pois representam os relacionamentos entre as mesmas, como *isAbout*, entre *Project* e *Topic.*

As propriedades de dados são também apresentadas como o rótulo das setas, mas agora entre uma classe e um tipo de dado. É compreendida como uma característica da instância de

uma classe, como ocorre com a propriedade *hasName*, que refere-se a classe *Person* e o tipo de dados *string*.

Há ainda outros tipos de propriedade, que não estão apresentadas no diagrama, como a propriedade de anotação que, como o nome sugere, serve para guardar informações sobre uma entidade ou outras propriedades.

Os tipos de dados (*Datatype)* são apresentados pelos retângulos. Eles especificam a que tipo de dado as instâncias de determinado conceito corresponderão. No exemplo temos que a classe *Person* tem um nome *hasName* e esse nome será do tipo *string.*

As instâncias, embora não estejam apresentadas na Figura 3, são os objetos das classes. Uma instância da classe *Person* pode ser, por exemplo, o nome de uma pessoa.

A formalidade das ontologias leva a compreensão de que a sua estrutura é baseada em lógica. Nesse sentido, as lógicas de descrição (DL) têm sido amplamente utilizadas para o desenvolvimento de ontologias. Essas lógicas permitem a especificação de classes e vários operadores, tais como união e intersecção [71]. A presença desses operadores lógicos dá o nome e caracteriza a DL e consequentemente dá expressividade às ontologias.

2.4 Integração de dados baseada em ontologias

Num processo de integração de dados existem diferentes etapas e as ontologias podem ser utilizadas em cada uma delas. Seja no acesso aos dados, nas suas caracterizações ou na visão global, o uso das ontologias tem se tornado cada vez mais frequente. Nessa seção serão apresentadas algumas metodologias e estudos de caso de integração de dados que utilizam ontologias. À solução que faz uso de ontologias para promover uma visão unificada dos dados integrados é dado o nome de *Ontology-Based Data Integration* (OBDI).

Na solução OBDI há a presença de uma ontologia que descreve o domínio de interesse e o conecta com um conjunto de dados, geralmente um conjunto já estruturado e existente. No entanto, as abordagens diferenciam-se quando a OBDI apresenta uma descrição semântica sobre o domínio de interesse, bem como o relacionamento entre os conceitos, permitindo assim uma remodelação dos dados sem mexer na estrutura real deles, armazenada nas bases de dados. Como resultado há uma independência entre o que é conceptualmente modelado e os dados [71], [72]. Desse modo, um aspeto importante numa solução OBDI é o facto de uma mesma entidade conceptual poder ser armazenada em bases de dados diferentes, sendo representadas por identificadores distintos [73].

Segundo os autores em [71], uma visão mínima de solução de OBDI possui três componentes essenciais:

1) A(s) ontologia(s), podendo existir uma ou mais camadas de ontologias e cada uma dela poder conter uma ou mais ontologias;
2) A(s) fonte(s) de dado(s), que geralmente são numerosas, heterógeneas e independentes, mas que pode ser uma só, passando a nomear a solução de *Ontology- Based Data Access* (OBDA), já que não há fontes de dados para integrar; e
3) Os mapeamentos entre os dois componentes anteriores, que são especificações precisas da correspondência entre os dados e os conceitos presentes na(s) ontologia(s).

Um pouco mais complexa, o autor em [72] apresenta uma perspetiva que envolve um *software* como uma componente de interrelação do sistema, como pode ser visto na Figura 4. Os autores argumentam que um sistema OBDI contém pelo menos três camadas, representadas pelas caixas maiores, com traço contínuo:

1) As fontes de dados;
2) As ontologias;
3) A aplicação de *software* que se utilizará na solução.

Os mesmos autores também alegam que existem variáveis a serem consideradas na hora de desenvolver a solução sendo representadas dentro das caixas pontilhadas da Figura 4.

Essas variáveis são:

1) Aquisição e acesso de fontes de dados: A aquisição dos dados e o acesso a eles vai depender do formato em que as fontes de dados estão descritas e o nível de acesso fornecido a elas. Geralmente as fontes de dados estão em formato estruturado, mas nem sempre estarão disponíveis numa base de dados para que o acesso ocorra através deles;
2) Mapeamento e transformação entre ontologias: Caso o programador da solução opte por uma abordagem com mais de uma camada de ontologias, o mapeamento e a transformação dos conceitos entre as camadas de ontologias deverá ocorrer de modo que a comunicação entre as camadas seja atingida. Para isso, podem ser utilizadas linguagens de mapeamento e *frameworks*[27] que auxiliam no mapeamento dos conceitos;

[27] https://pt.wikipedia.org/wiki/Framework [Consult. em 23/08/2021].

3) Armazenamento do conteúdo da ontologia e dos dados: O conteúdo da ontologia pode ser armazenado numa base de dados própria para ontologias (RDF *Triple Store*)[28], localmente (com a materialização ou não dos dados) ou em bases de dados relacionais com acesso lógico aos dados por meio de soluções existentes no mercado;
4) Linguagem e *framework* para a construção das ontologias: Geralmente a linguagem utilizada para a construção das ontologias é o RDF/RDFS em conjunto com o OWL;
5) Acesso à solução de ODBI: O resultado do sistema deverá ser acedido por outras soluções ou pelo utilizador final. Usualmente a resposta do sistema é o resultado das consultas executadas.

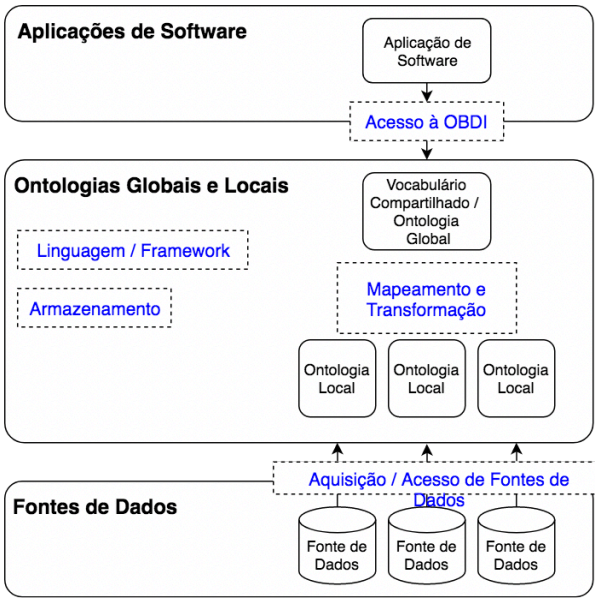

Figura 7: Representação dos elementos de uma solução OBDI «Fonte: autores em [74]»

2.5 Abordagens de Ontologias

O instanciamento de dados numa ontologia pode ser concretizado de diferentes formas. A isto os autores em [74] chamam de abordagens, sendo que, das abordagens clássicas, as mais utilizadas são: a abordagem de ontologia única, a abordagem de múltiplas ontologias, a abordagem híbrida e a abordagem GAV, explicadas a seguir.

1) **Abordagem de Ontologia Única**: Uma única ontologia global é construída. Desse modo o processo de desenvolvimento consiste em definir a ontologia global e realizar os mapeamentos das fontes de dados para a ontologia. Essa abordagem é a mais simples e rápida, porém é a mais difícil de manter, principalmente quando há adição ou

[28] https://en.wikipedia.org/wiki/Triplestore [Consult. em 23/08/2021].

atualização de fontes de dados [74]. Ainda sim, esse tipo de abordagem é recomendada quando todas as fontes de dados são estáveis e possuem a mesma visão do domínio [69];

2) **Abordagem de Múltiplas Ontologias**: Aqui são construídas apenas ontologias locais das respetivas fontes de dados, os mapeamentos entre as fontes de dados e as ontologias e os mapeamentos entre as ontologias. O mapeamento busca encontrar conceitos similares entre as ontologias, o que pode não ser uma tarefa simples, dependendo muito do especialista do domínio. Outro ponto a ser ponderado é que a cada nova fonte de dados novos mapeamentos entre ontologias devem ser feitos [74].
3) **Abordagem Híbrida**: Nessa abordagem utilizam-se os mecanismos das duas abordagens anteriores, onde há o desenvolvimento de ontologias locais para cada uma das fontes de dados e um vocabulário compartilhado, que pode ser também uma ontologia global. Desse modo, há inicialmente a construção do vocabulário compartilhado e a partir dele as ontologias locais são construídas, auxiliando os mapeamentos entre as ontologias. Logo a seguir, o conteúdo das ontologias locais é então mapeado para as fontes de dados [74]. O desenvolvimento dessa abordagem é mais complexo, porém possui a vantagem na facilidade de adição de fontes de dados, uma vez que o vocabulário comum não sofrerá alterações, sendo necessária apenas o acréscimo dos mapeamentos relativos a nova fonte [69].
4) **Abordagem GAV**: Nessa abordagem segundo os autores em [74], são desenvolvidas as ontologias globais e as locais, entretanto o acesso ao sistema dá-se unicamente através da ontologia global. O primeiro passo no desenvolvimento do sistema é a determinação do vocabulário das ontologias locais, seguido pela realização dos mapeamentos dessas ontologias com as fontes de dados. Depois é desenvolvida a ontologia global que receberá o mapeamento do vocabulário das ontologias locais. Essa abordagem é indicada quando há a presença de ontologias existentes ou em uso e existe a necessidade de preservar a estrutura das mesmas.

Na Figura 8 há uma esquematização dessas abordagens criada pelos autores em [74]. Nela são apresentadas quatro camadas [A] das fontes de dados, [B] relativa às ontologias locais, [C] a camada da ontologia global ou vocabulário compartilhado e [D] a camada das aplicações de *software* que terão acesso à solução. As setas pontilhadas indicam o acesso aos dados por meio das aplicações. As setas contínuas representam a transformação ou a virtualização do acesso aos dados. As setas tracejadas representam as relações implícitas de herança entre as

ontologias. Já as elipses com os números representam a sequência de desenvolvimento das soluções.

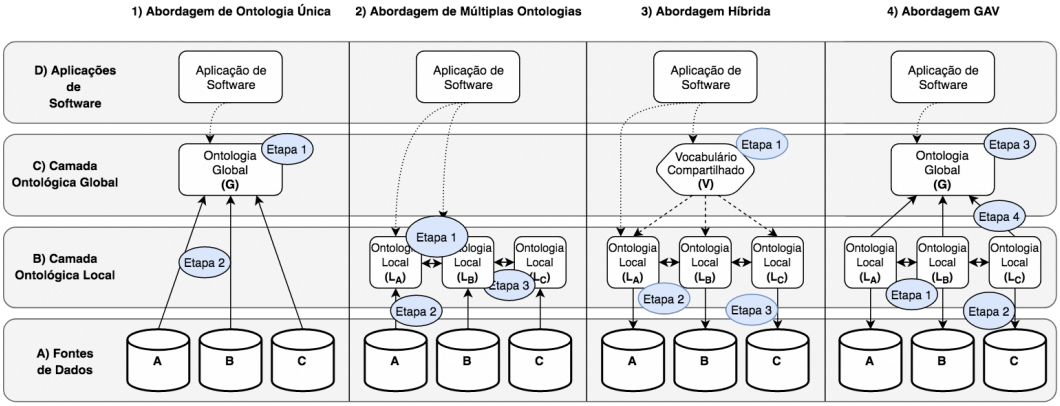

Figura 8: Abordagens de OBDI «Fonte: autor em [61]»

3 Estado da arte da *web* semântica

A aplicabilidade da *web* semântica é muito atrativa quando se quer resolver um problema de pesquisa que forneça informações de valor para tomada de decisões, principalmente porque fornece a possibilidade destas decisões serem tomadas por máquinas e não só por pessoas. Abre-se então um leque de exploração jamais visto. Nesse quesito, muitos trabalhos já existem e alguns deles serão mencionados a seguir segundo a sua importância no contexto desta dissertação e, principalmente, por apresentarem soluções ou propostas de soluções diferentes em termos de área de estudo, porém com similaridade quanto ao assunto relacionado à *web* semântica.

3.1 *A web* semântica

Zaidan e Bax em [77] explicam que os dados abertos vinculados *Linked Open Data* (LOD) têm potencial para melhorar a qualidade dos dados dos sistemas de informação nas mais diversas áreas. Uma das áreas é a gestão da informação clínica. No artigo em questão os autores apresentam o LOD, a fim de constatar a possibilidade de se agregar valor aos dados de sistemas de informação clínica.

O método que utilizaram foi propor uma pesquisa sobre a *web* semântica na qual efetivou-se uma revisão de literatura dos principais conceitos. Uma pesquisa documental foi realizada obtendo uma bibliografia baseada em livros e artigos de autores conceituados. Como resultado, foram apresentadas seis tecnologias que utilizam o padrão de dados abertos vinculados e que propõem a agregação de valor aos dados de sistemas de informação, em especial aos dados clínicos. Os pontos em comum destas tecnologias situam-se na publicação, extração, interligação e consumo dos dados do LOD.

Eles concluíram que é possível agregar valor aos dados internos dos sistemas de informação clínica, mesmo quando se tem disponível um número enorme de triplos do LOD para serem utilizados e/ou consumidos. A continuidade dá-se na medida em que se faça uma aplicação prática do mesmo, ou seja, utilizando uma ou mais das tecnologias exemplificadas no artigo, será portanto factível de comprovar que o valor que se pode agregar ao utilizar os triplos RDF do LOD no domínio da saúde, aos dados internos dos sistemas de informação clínicos, proporcionando melhores informações aos interessados.

O conteúdo do artigo foi de fundamental importância na construção deste trabalho pois conseguiu contribuir para a estruturação teórica e, além disto, a similaridade da proposta de criar uma base para futura exploração foi fator decisivo para inspiração do mesmo.

3.2 Potencialidades e tendências na nova geração de ambientes de ensino na *internet*

Os autores em [78] exploram a área da educação. Eles afirmam que as tecnologias educacionais com base na *web* têm obtido excelentes resultados nas últimas décadas, principalmente por causa dos diversos cursos oferecidos na modalidade "educação à distância" onde a *Internet* é a plataforma base de comunicação e interação entre alunos e professores. Desta forma, segundo eles, duas linhas de pesquisa estão em crescente expansão.

A primeira delas é a *web s*emântica que desenvolve tecnologias que permitem ao computador compartilhar e manipular as informações contidas na *web* de forma adequada e inteligente. Com esta tecnologia em ambientes de aprendizagem, os agentes de *software* podem interagir entre si, trocar informações e auxiliar professores e alunos a selecionar, combinar e classificar o conteúdo disponível na *web*.

E a segunda é a *web* 2.0, ou *web* social, onde os utilizadores são beneficiados por diversas ferramentas para compartilhar e construir "conhecimento" de forma simples, interativa e colaborativa, tais como o *wiki*[29], que é um *website* que pode ser lido e editado por qualquer pessoa (devidamente autorizada) e sem a necessidade de ter nenhum conhecimento tecnológico avançado; os *websites* para partilha de fotos, vídeos e materiais educacionais; e por último, os *websites* de relacionamento, que tem a maior popularidade na *web* 2.0. Estes *websites*, como o *Facebook*[30], permitem que os utilizadores mantenham contacto com os seus amigos e familiares, além de viabilizar a partilha de diferentes tipos de conteúdo e a criação de aplicativos (por qualquer utilizador) que podem ser inseridos diretamente na página principal.

Segundo os autores, esta recente interseção destas tecnologias, à altura do artigo, iriam promover novas formas de aprendizado envolvendo alunos e professores e resolver os diversos problemas apresentados, através de propostas de soluções que possivelmente teriam impacto muito positivo na qualidade do ensino oferecido pela *web*.

[29] https://en.wikipedia.org/wiki/Main_Page [Consult. em 23/08/2021].
[30] https://pt-pt.facebook.com/ [Consult. em 23/08/2021].

Da mesma forma, a proposta desta dissertação é promover e difundir a tecnologia da *web* semântica de forma a auxiliar estudantes na tomada de decisões. Espera-se porém, que, a partir deste trabalho e para além dos limites do mesmo, haja continuidade nas investigações de modo a que possa ser explorada a inserção na *social web* ou ainda o uso de tecnologias envolvendo inteligência artificial (IA) em conjunto com a *web* semântica.

3.3 A aplicação da *web* semântica no jornalismo

O texto apresentado pelos autores Lammel e colegas [79] mostra que a *web* e as bases de dados são consideradas plataformas tecnológicas fundamentais para o desenvolvimento do jornalismo contemporâneo em redes digitais, principalmente depois do surgimento da *web* semântica nos anos 2000 como proposta de expansão da atual *web*, para torná-la mais automatizada e eficiente.

Um exemplo disto é o BBC Wildlife[31], um *website* que utiliza tecnologias da *web* semântica para gerir e publicar conteúdos editoriais sobre o mundo natural, em que são analisados aspetos que contribuem na potencialização de características do jornalismo estruturado em bases de dados. Cada página do *website* monta a sua estrutura de navegação automaticamente, de acordo com os tipos de relações que os conteúdos possuem. Dessa forma, o sistema cria automaticamente uma malha de páginas interligadas, rica em relacionamentos. As diferentes maneiras de se categorizar os conteúdos permite que o *website* formule e distribua pelas páginas internas várias listas de *links*, que convidam o utilizador a continuar a navegação pelo *website* de acordo com o contexto, como se as próprias páginas internas fossem um grande menu de navegação.

Como resultado tem-se uma interoperabilidade automatizada, que permite que diferentes *websites* (que estejam na lógica da *web* semântica) troquem entre si dados e informações de maneira automatizada. Isto é feito a partir de associações de conceitos definidos por vocabulários ou ontologias em comum e pela ativação da memória, que, no caso do BBC Wildlife, permitiu o desenvolvimento de um novo produto através da exploração de um grande repositório de vídeos produzidos anteriormente que, até então, estavam arquivados.

Em outras palavras, tanto a interoperabilidade de dados quanto a automatização das máquinas na identificação de significados e na geração de inferências permitiram que milhares de vídeos fossem associados a conteúdos produzidos na *web* e que fossem integrados num

[31] https://www.discoverwildlife.com/

mesmo sistema, gerando, assim, um novo produto. O caso demonstra que a *web* semântica é um facilitador no processo de construção de bases de conhecimento, que podem vir a ser exploradas pelos *websites* jornalísticos, tornando a tecnologia fomentadora de possíveis rupturas dos produtos jornalísticos tradicionais.

Essa proposta de construção de bases de conhecimento e ruptura dos meios tradicionais de difundir a informação são exploradas nesta dissertação. As informações de cursos e sua ligação ao mercado de trabalho disponibilizadas em formato de dados abertos irá permitir às faculdades e instituições a criação de ferramentas que podem acrescentar mais valor às organizações académicas, incluindo a componente de seleção dos cursos de ensino superior por parte dos alunos.

3.4 SemanticSefaz

Rolim e colegas em [80] fizeram um trabalho para a Secretaria da Fazenda do estado do Ceará, no Brasil. Eles consideraram que a fiscalização no processo de contratação pública é considerada fundamental para a sociedade como meio de promover maior segurança e controlo contra possíveis fraudes e ações ilegais. No entanto, os dados disponíveis sobre as compras governamentais por si só não permitiam a identificação de possíveis contratos firmados ou licitações vencidas por empresas inaptas ou suspensas, dificultando a análise e fiscalização por funcionários governamentais. Além disso, muitas vezes os dados não estavam disponíveis no mesmo formato comum e diferenciavam no seu vocabulário, tornando difícil para esses profissionais encontrar informações relevantes. Como forma de solucionar esses problemas, fizeram o *SemanticSefaz*, um portal semântico de integração entre bases heterogéneas voltado para o domínio da contratação pública por meio de uma visão homogénea, permitindo consultas relacionadas a contratos e compras governamentais por parte de empresas sancionadas.

Como estudo de caso, as bases de dados com dados sobre compras governamentais, empresas insalubres, empresas em suspensão e empresas punidas, foram utilizadas para construir o portal semântico. Posteriormente, foram realizadas consultas de interesse do domínio tributário por meio do *SemanticSefaz*, demonstrando a sua eficiência na realização de consultas semânticas contribuindo para o objetivo principal de ser uma ferramenta para a integração, visualização e descoberta de conhecimentos que facilitam o trabalho dos profissionais tributários.

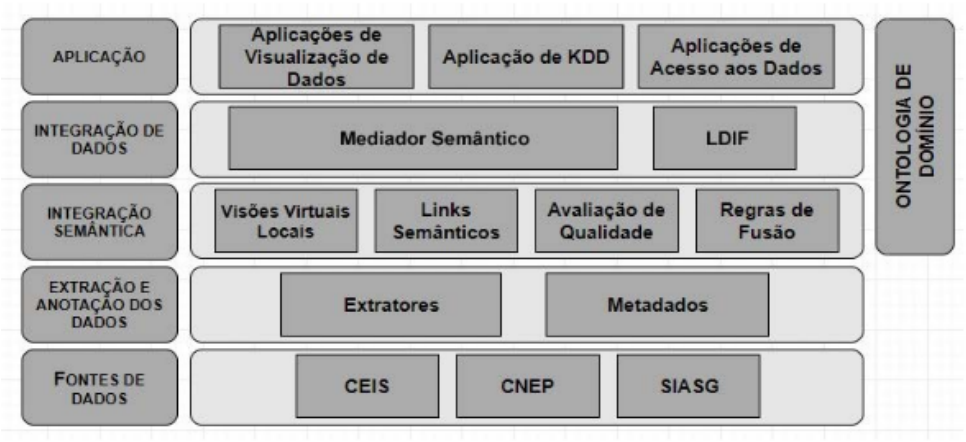

Figura 9: Arquitetura do SemanticSefaz. «Fonte: autores em [80]»

Segundo os autores, a arquitetura do *SemanticSefaz* é demonstrada na Figura 9 e cada uma das camadas é explicada a seguir, conforme descrito em [80]:

- **Fontes de dados**: Nesta camada estão as fontes de dados disponibilizadas pelo portal. As fontes de dados podem ser de diferentes tipos e proveniência.

- **Extração e anotação de dados**: Nesta camada é realizada a extração e anotação das fontes de dados disponíveis no portal. As anotações são metadados que descrevem desde o formato da fonte até comentários sobre os significados de elementos do esquema utilizado pela fonte.

- **Ontologia de Domínio**: Essa ontologia é usada como um meio para estabelecer um vocabulário formal e explícito a ser compartilhado para a anotação semântica das fontes de dados.

- **Integração Semântica (IS)**: O resultado da IS define uma visão semântica integrada das múltiplas fontes de dados.

- **Integração de Dados**: O Portal disponibiliza dois serviços de acesso às fontes de dados locais através da visão ontológica.

- **Mediador Semântico:** Permite que os utilizadores possam definir consultas ad-hoc sobre as ontologias de domínio em SPARQL, reformulando a consulta em consultas sobre as fontes de dados nas suas linguagens de consultas nativas.

- **Triplificador:** Gera uma representação materializada em RDF da visão semântica. Este módulo realiza as etapas de limpeza e resolução de conflitos, garantindo a qualidade do resultado.

- **Aplicação**: Na última camada estão os utilizadores e aplicações que fazem uso dos serviços de integração dos dados.

O *SemanticSefaz* foi apresentado por Rolim *et al.* em [80] juntamente com seus componentes, bem como seu *workflow*, explicitando o passo a passo para realização de consultas em cima das fontes de dados através das vistas semânticas. Caracteriza-se como um portal semântico tendo como base uma abordagem virtualizada de mediação em vistas semânticas para fornecimento de uma integração entre bases do domínio de fiscalização de gastos públicos. Ele foi concebido de tal forma que facilita o trabalho de profissionais fiscais de maneira prática e sem necessitar de conhecimentos acerca de *web* semântica, *linked data* ou das linguagens de consultas SQL/SPARQL.

Nesta dissertação, pretende-se aproveitar a proposta do *SemanticSefaz* na construção da arquitetura e na sua aplicabilidade de forma a criar um portal de disponibilização de informações com acesso livre e com dados abertos, que facilitará a tomada de decisão de qualquer pessoa que necessite escolher um curso que definirá a sua carreira profissional.

A metodologia deste trabalho foi feita tendo por base a metodologia aplicada na construção do *SemanticSefaz*. Especificamente foram aproveitados os conceitos de fontes e extração de dados, ontologias, integração de dados e apresentação no portal de um exemplo de aplicação.

Da mesma forma que no *SemanticSefaz* há o conceito de fontes e extração de dados, bem como o armazenamento destes em base de dados local, entretanto, de forma a facilitar a integração semântica e instanciamento dos dados na ontologia, as tabelas dessas bases de dados foram construídas tendo por base a ontologia, entretanto, o que difere do *SemanticSefaz* é que este utiliza uma ontologia intermediária para permitir a integração semântica e posterior a integração dos dados enquanto que neste trabalho não foi utilizada nenhuma ontologia intermediária, apenas uma única ontologia de domínio que abrange o mapeamento e a integração semântica de forma consistente.

Quanto à aplicação, no *SemanticSefaz* há alguma complexidade na sua utilização regular, o que torna obrigatória e necessária manutenção incremental das bases de dados envolvendo várias fontes de dados de forma sistemática, deixando a cargo das entidades interessadas a tarefa de fornecer as informações necessárias para a integração ao portal de forma estruturada. No presente trabalho, optou-se por utilizar uma única ontologia e várias fonte de dados limitando seu propósito no uso académico, portanto, a manutenção incremental é

prevista, porém não considerada. Obviamente, em projetos para além do uso académico, será necessária a consideração deste aspeto da arquitetura do *SemanticSefaz* para garantir a longevidade do mesmo.

4 Metodologia

Esse Capítulo detalha a metodologia de desenvolvimento deste trabalho.

Para atingir os objetivos propostos, foram seguidos os passos enumerados no Capítulo 1 (seção 1.2), culminado com o desenvolvimento de uma aplicação *web*, chamada Portal Semantic-ESP (PSESP)[32]. Pressupõe-se que o PSESP tenha como requisitos, a capacidade de extrair dados na *web* em diversos formatos, armazenar esses dados numa base de dados local, instanciar esses dados na ontologia e disponibilizar as informações instanciadas na *web* em formato aberto e público. Seguindo esta linha de trabalho, o desenvolvimento foi dividido em tarefas, que são assim enumeradas:

1. Criação da ontologia OP
2. Criação das Tabelas Relacionais (TR)
3. Criação do Mapeamento das Tabelas Relacionais e a ontologia
4. Extração dos dados de cursos (DC) e dados de profissões (DP)
5. Instanciamento da ontologia OP
6. Apresentação dos dados em formato de dados abertos

As considerações sobre cada uma dessas tarefas serão explicadas detalhadamente nas seções seguintes.

4.1 Criação da ontologia OP

Esta primeira tarefa consiste na criação da ontologia OP. Como foi visto no Capítulo 2 (seção 2.5), existem várias abordagens na construção de qualquer ontologia. Dentro das mencionadas, a abordagem que melhor se ajusta ao desenvolvimento deste trabalho é a abordagem de ontologia única. Assim, no presente trabalho, foi construída a ontologia OP que engloba os dados das instituições e respetivos cursos além dos dados de saídas profissionais.

A escolha desta abordagem deve-se ao facto de simplificar o mapeamento entre as fontes de dados. Para a instanciação da ontologia OP foram consideradas múltiplas fontes de dados, entretanto novas fontes de dados podem ser utilizadas no futuro com pequenas alterações na ferramenta. As novas fontes de dados devem ser codificadas, uma vez que parte das

[32] https://github.com/LuSoMaBra/semantic-esp

informações dos cursos são extraídas de *websites* de instituições no formato HTML e cada um dos *websites* tem um código diferente.

A criação da ontologia foi baseada no trabalho do Professor Ibrahim *et al.* em [81] que trata sobre recomendação de cursos utilizando uma abordagem híbrida baseada em ontologias. Na referida tese, foi feita a criação de três ontologias, a saber, *Course Ontology*, que faz referência aos cursos de universidades, *Job Ontology*, que referencia as saídas profissionais e *User Ontology*, que faz referência aos utilizadores que receberão as recomendações de cursos após o tratamento da solução proposta. Com respeito a ontologia *User Ontology* nada se aproveitou por fugir do âmbito deste trabalho, entretanto, os princípios de reuso dos vocabulários das ontologias *Course Ontology* e *Job Ontology* serviram de inspiração para criação da ontologia OP, cujo desenvolvimento e mais detalhes serão demonstrados no Capítulo 5.

4.2 Criação das Tabelas Relacionais

Após a criação do vocabulário da ontologia OP (tarefa 1), a próxima tarefa é a criação de uma estrutura de base de dados local para armazenamento dos DC e DP. Nessa tarefa, para facilitar o mapeamento entre a base de dados e a ontologia, procurou-se sempre que possível, utilizar os mesmos vocabulários. Assim, os nomes das tabelas são os nomes das classes e os nomes dos atributos são os mesmos nomes das propriedades dos objetos. Assim, os mapeamentos entre a ontologia e a base de dados são feitos diretamente sem a necessidade de uma normalização intermediária.

Mais detalhes sobre a criação das Tabelas Relacionais (TR) serão apresentados no Capítulo 6 (seção 6.1.1).

4.3 Criação do Mapeamento das Tabelas Relacionais e a Ontologia

Para criar o mapeamento de *Relational Database* (RDB) para RDF, os autores em [82], [83] e [84] mostram que o Grupo de Trabalho W3C RDB2RDF[33] recomenda duas abordagens: Mapeamento Direto (MD) e R2RML[34]. O MD fornece um conjunto de regras de mapeamento de acordo com o esquema RDB, enquanto o R2RML permite aos utilizadores definir manualmente os mapeamentos de acordo com a ontologia destino. O principal problema de usar

[33] https://www.w3.org/2001/sw/rdb2rdf/ [Consult. em 23/08/2021].
[34] https://www.w3.org/TR/r2rml/ [Consult. em 23/08/2021].

R2RML é o esforço para criar um mapeamento de documentos R2RML manualmente. Isso pode levar a muitos erros no documento R2RML e requer domínio de especialistas.

Para o mapeamento entre a base de dados e a ontologia OP optou-se pela abordagem MD. Para tal, existem várias ferramentas especializadas em executar a tarefa de mapear RDB para RDF. Dentro destas, podem ser citados "R2RML *By Assertions*"[35], RDB2RDF *Plugin for Eclipse*[36], Virtuoso RDF *View*[37] e outros, entretanto optou-se por implementar o mapeamento diretamente dentro do código, o qual é explicado em detalhes no Capítulo 6 (seção 6.2).

Como mostrado na seção 4.2, a construção das TR foi baseada no vocabulário da ontologia OP. Assim, o mapeamento é feito associando o nome de um objeto da ontologia OP ao objeto correspondente da TR, o qual tem o mesmo nome do objeto da ontologia. Isto é feito através de uma tabela de associação. A criação deste mapeamento é descrita em detalhes no Capítulo 6 (seção 6.2).

4.4 Extração dos dados DC e DP

O instanciamento da ontologia só será possível se houver dados disponíveis. Para esta tarefa, foi construído um módulo específico de *software* para extrair os dados do portal de dados abertos do governo e do portal de saídas profissionais, especificados no Capítulo 1. Este módulo extrai as informações no formato JSON e no formato HTML, conforme o caso.

A extração de dados DC e DP é descrita em detalhes no Capítulo 6 (seção 6.1).

4.5 Instanciamento da Ontologia OP

Para o instanciamento da OP foi utilizado um código interno especificamente criado para isso que utiliza o mapeamento definido após a realização da tarefa 3. Na presente tarefa, os dados são serializados em formato *Java Script Object Notation* (JSON)[38], para que seja possível a disponibilização em formato aberto. Este formato foi escolhido por defeito em detrimento de outros, tais como YAML[39] ou XML[40], pela simplicidade de manuseio e entendimento, além de maior disponibilidade de bibliotecas para as ferramentas escolhidas. É

[35] https://www.researchgate.net [Consult. em 23/08/2021].
[36] https://www.eclipse.org/eclipseide/ [Consult. em 23/08/2021].
[37] https://virtuoso.openlinksw.com/ [Consult. em 23/08/2021].
[38] https://en.wikipedia.org/wiki/JSON [Consult. em 23/08/2021].
[39] https://en.wikipedia.org/wiki/YAML [Consult. em 23/08/2021].
[40] https://en.wikipedia.org/wiki/XML [Consult. em 23/08/2021].

importante frisar que o PSESP permite os acessos aos dados noutros formatos, conforme a escolha do utilizador.

Mais detalhes sobre o instanciamento da OP são descritos no Capítulo 6 (seção 6.2).

4.6 Apresentação dos dados em formato de dados abertos

Para a apresentação dos dados recolhidos, foi criado o portal PSESP, com acesso público, que permite o acesso às informações de forma estruturada e em formato de dados abertos no formato JSON, mas contendo opções de outros formatos conforme a escolha do utilizador. O PSESP oferece também ao utilizador a possibilidade de escrever uma consulta SPARQL e executá-la sobre os dados apresentados, permitindo uma interação direta e mais técnica aos dados por parte do utilizador.

Mais detalhes sobre a apresentação dos dados em formato aberto são descritos no Capítulo 6 (seção 6.3).

5 Desenvolvimento da Ontologia OP

Como visto no Capítulo 2 (seção 2.4), o desenvolvimento de qualquer ontologia deve ser orientado pelos seguintes passos:

1 Determinar o âmbito;
2 Considerar o reuso de outra ontologia ou termos;
3 Enumerar os termos;
4 Definir as classes;
5 Definir as propriedades;
6 Definir as restrições ou axiomas;
7 Criar as instâncias.

Na determinação do âmbito é definido o domínio do qual as informações serão estruturadas, assim como as perguntas que se deseja serem respondidas com essa ontologia. Neste passo também se faz uma aquisição do conhecimento do domínio, enumerando os principais termos e a partir desses termos, definem-se as classes da ontologia, as propriedades de objetos e dados, definem-se os axiomas, se existirem, e por fim criam-se as instâncias.

Para a criação e mapeamento da ontologia foi utilizada a ferramenta Protegé[41]. A escolha desta ferramenta foi feita pela flexibilidade e possibilidade de exportação em vários formatos. Para apresentação gráfica, foi utilizada a ferramenta WebVOWL[42] que permite a importação da ontologia criada pelo Protegé e a apresentação automática em formato visualmente atrativo.

Na criação da ontologia OP, o âmbito foi dividido em duas partes separadas, porém importantes para a formulação da OP. A primeira parte diz respeito aos cursos superiores de universidades e institutos portugueses cujo objetivo é recuperar informações sobre os cursos superiores de universidades e institutos portugueses tais como nome dos cursos, faculdades, e outras informações relevantes. A segunda parte diz respeito às ofertas profissionais de trabalho em Portugal cujo o objetivo é recuperar informações sobre as ofertas de empregos atualizadas e disponíveis em território português, incluindo os requisitos educacionais para tal oferta. Esses requisitos são importantes pois serão utilizados na fase de integração.

[41] https://protege.stanford.edu/ [Consult. em 23/08/2021].
[42] http://www.visualdataweb.de/webvowl/ [Consult. em 23/08/2021].

Com relação ao reuso de outra ontologia ou termos, embora não tenha sido encontrada uma ontologia completa que fosse adequada à necessidade desse trabalho, foi possível reusar a maioria dos termos os quais foram buscados dos vocabulários das ontologias: schema[43], dcterms[44], owl[45], rdf[46], rdfs[47], xml[48], xsd[49] e dbo[50].

O passo seguinte no desenvolvimento de uma ontologia é a enumeração dos termos. Os termos considerados relevantes para a OP são especificados abaixo, juntamente com os prefixos utilizados, onde o prefixo "psesp" corresponde aos vocabulários e termos criados para a ontologia OP:

Prefixos:

1) owl: <http://www.w3.org/2002/07/owl#>
2) rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3) rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4) xml: <http://www.w3.org/XML/1998/namespace>
5) xsd: <http://www.w3.org/2001/XMLSchema#>
6) dcterms: <http://purl.org/dc/terms/>
7) schema: <http://schema.org/>
8) psesp:<ontologia_psesp.owl#>

Termos:

1) schema:EducationalOcuppationalProgram
9) schema:CollegeOrUniversity
10) dcterms:ProvenanceStatement
11) schema:JobPosting
12) schema:PostalAddress
13) schema:educationRequirements
14) schema:provider
15) schema:address

[43] https://schema.org/ [Consult. em 23/08/2021].
[44] http://purl.org/dc/terms/> [Consult. em 23/08/2021].
[45] http://www.w3.org/2002/07/owl
[46] http://www.w3.org/1999/02/22-rdf-syntax-ns
[47] http://www.w3.org/2000/01/rdf-schema
[48] http://www.w3.org/XML/1998/namespace
[49] http://www.w3.org/2001/XMLSchema
[50] http://dbpedia.org/ontology/

16) schema:name
17) schema:description
18) schema:educationalOccupationalProgram
19) schema:title
20) schema:branchCode
21) schema:url
22) schema:streetAddress
23) schema:addressLocality
24) schema:addressRegion
25) schema:postalCode
26) dcterms:creator
27) dcterms:created
28) dcterms:modified
29) psesp:cnaefArea
30) psesp:programmeArea
31) schema:termDuration
32) schema:educationalCredentialAwarded
33) psesp:internationalRegistrationFee
34) psesp:nationalRegistrationFee
35) schema:qualifications
36) schema:baseSalary
37) schema:employmentType
38) psesp:lastExtraction
39) schema:fileFormat

A descrição de cada termo da ontologia OP são apresentados nas Tabelas 2, 3 e 4. A Tabela 2 apresenta a definição das classes da OP. A coluna “Descrição da classe” desta tabela mostra a descrição da classe bem como o url de origem do termo, quando aplicável.

Tabela 2 – Tabela de Classes da OP «Fonte: elaboração própria»

Nome da classe	Descrição da classe
schema:EducationalOcuppationalProgram	Dados dos cursos superiores incluindo a instituição a que pertence. *website:* https://schema.org/EducationalOccupationalProgram
schema:CollegeOrUniversity	Dados das instituições de ensino superior. *website:* https://schema.org/CollegeOrUniversity
dcterms:ProvenanceStatement	Dados do provedor da informação tais como, origem e data de atualização e extração. *website:* https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/ProvenanceStatement/
schema:JobPosting	Dados das ofertas de trabalho. *website:* https://schema.org/JobPosting
schema:PostalAddress	Morada das instituições de ensino superior. *website:* https://schema.org/PostalAddress

A tabela apresenta as propriedades de objetos da ontologia OP. A coluna "Descrição" tem a mesma função da tabela 2. A coluna "Domínio" mostra o domínio ao qual pertence a propriedade e a coluna "Alcance" mostra o alcance ou "*range*" que a referida propriedade abrange.

Tabela 3 – Tabela de propriedades de objetos da OP «Idem»

Propriedade	Domínio	Alcance	Descrição
schema:provider	schema:EducationalOcuppationalProgram	schema:CollegeOrUniversity	Define qual curso é disponibilizado por qual instituição. *website:* https://schema.org/provider
schema:educationRequirements	schema:JobPosting	schema:EducationalOcuppationalProgram	Define a ligação entre os cursos e as saídas profissionais através da exigência de formação superior. *website:* https://schema.org/educationRequirements
schema:provider	schema:CollegeOrUniversity, EducationalOcuppationalProgram, JobPosting	schema:ProvenanceStatement	Define o registo de acesso e origem de dados para cursos, instituições e ofertas de trabalho. *website:* https://schema.org/provider
schema:address	schema:CollegeOrUniversity	schema:PostalAddress	Define o registo de moradas das instituições. *website:* https://schema.org/address

A tabela 4 apresenta as propriedades de dados da ontologia OP. As colunas "Domínio" e "Descrição" tem a mesma função da tabela 3, sendo que quando o termo é específico para a ontologia OP também são apresentados alguns exemplos. A coluna "Tipo" apresenta a estrutura de dados associados à propriedade.

Tabela 4 – Tabela de propriedades de dados da OP «Idem»

Propriedade	Domínio	Tipo	Descrição
schema:name	schema:EducationalOcuppationalProgram, schema:CollegeOrUniversity	*string*	Nome do curso ou da instituição. *website*: https://schema.org/name
schema:description	schema:EducationalOcuppationalProgram, schema:JobPosting	*string*	Descrição do curso ou da oferta de trabalho. *website*: https://schema.org/description
schema:educationalProgramMode	schema:EducationalOcuppationalProgram	*string*	Descrição do modo do curso (ex: diurno, noturno, online, etc.). *website*: https://schema.org/educationalProgramMode
schema:title	schema:JobPosting, schema:ProvenanceStatement	*string*	Título da oferta de trabalho ou do provedor de dados. *website*: https://schema.org/title
schema:branchCode	schema:CollegeOrUniversity	*string*	Código nacional da instituição. *website*: https://schema.org/branchCode
schema:url	schema:EducationalOcuppationalProgram, schema:CollegeOrUniversity, schema:ProvenanceStatement	*string*	Url da instituição ou do curso ou do provedor de dados. *website*: https://schema.org/url
schema:streetAddress	schema:PostalAddress	*string*	Nome da rua associada a morada da instituição. *website*: https://schema.org/streetAddress
schema:addressLocality	schema:PostalAddress	*string*	Nome da cidade associada a morada da instituição. *website*: https://schema.org/addressLocality
schema:addressRegion	schema:PostalAddress	*string*	Nome do distrito associado a morada da instituição. *website*: https://schema.org/addressRegion
schema:postalCode	schema:PostalAddress	*string*	Nome do código postal associado a morada da instituição. *website*: https://schema.org/postalCode
dcterms:creator	schema:ProvenanceStatement	*string*	Nome do criador da informação. *website*: https://dublincore.org/specifications/dublin-core/dcmi-terms/
dcterms:created	schema:ProvenanceStatement	*datetime*	Data da criação da informação. *website*: https://dublincore.org/specifications/dublin-core/dcmi-terms/
dcterms:modified	schema:ProvenanceStatement	*datetime*	Data da última atualização da informação. *website*: https://dublincore.org/specifications/dublin-core/dcmi-terms/
schema:termDuration	schema:EducationalOcuppationalProgram	*string*	Duração do curso em semestres. *website*: https://schema.org/termDuration

schema:educationalCredentialAwarded	schema:EducationalOcuppationalProgram	*string*	Nível da formação que se espera ao final do curso (ex. Mestrado, Licenciatura, etc.) *website*: https://schema.org/educationalCredentialAwarded
schema:qualifications	schema:JobPosting	*string*	Define a qualificação requerida para ocupação da vaga oferecida na oferta de trabalho. *website*: https://schema.org/qualifications
schema:baseSalary	schema:JobPosting	*string*	Define o valor do salário oferecido para a vaga da oferta de trabalho. *website*: https://schema.org/baseSalary
schema:employmentType	schema:JobPosting	*string*	Define o tipo da oferta de trabalho (ex.: tempo integral, teletrabalho, etc.). *website*: https://schema.org/employmentType
schema:fileFormat	schema:ProvenanceStatement	*string*	Define o formato da origem dos dados (ex.: HTML, JSON, etc.). *website*: https://schema.org/fileFormat
psesp:programmeArea	schema:EducationalOcuppationalProgram	*string*	- Propriedade da ontologia OP; - Representa a área da Classificação Nacional de Áreas de Educação e Formação (CNAEF); - Valor esperado é uma *string* texto contendo o nome da área; - Ex.: “Ciências Agrárias”, “Contabilidade”, “Biotecnologia”, etc.
psesp:internationalRegistrationFee	schema:CollegeOrUniversity	*string*	- Propriedade da ontologia OP; - Representa o valor da propina anual para alunos internacionais; - Valor esperado é uma *string* texto contendo o valor anual da propina.
psesp:nationalRegistrationFee	schema:CollegeOrUniversity	string	- Propriedade da ontologia OP; - Representa o valor da propina anual para alunos nacionais; - Valor esperado é uma *string* texto contendo o valor anual da propina.
psesp:lastExtraction	schema:ProvenanceStatement	*datetime*	- Propriedade da ontologia OP; - Representa a data e hora da última extração efetuada; - Valor esperado é um atributo *datetime* contendo a data e hora da última extração.

Após definir as classes e propriedades, o passo a seguir no desenvolvimento de uma ontologia é a definição das restrições e axiomas. A tabela 5 apresenta as restrições, onde as colunas “Domínio” e “Alcance” tem a mesma função da tabela 3.

Tabela 5 – Tabela de Restrições da OP «Idem»

Restrição	Domínio	Alcance
schema:provider	schema:CollegeOrUniversity	schema:EducationalOcuppationalProgram
schema:educationRequirements	schema:JobPosting	schema:EducationalOcuppationalProgram
schema:provider	schema:ProvenanceStatement	schema:CollegeOrUniversity
schema:provider	schema:ProvenanceStatement	schema:EducationalOcuppationalProgram
schema:provider	schema:ProvenanceStatement	schema:JobPosting
schema:address	schema:PostalAddress	schema:CollegeOrUniversity

A última fase no desenvolvimento de uma ontologia é a criação de suas instâncias. O instanciamento dos dados na ontologia OP é feito à medida que o utilizador solicita acesso aos dados no portal PSESP. Isso será explicado em detalhes no Capítulo 6 (seção 6.2).

O leitor pode encontrar o código fonte da OP em sintaxe *RDF/XML*[51] no Anexo 04 – Código fonte da OP. A representação gráfica da OP[52] está demonstrada na Figura 10.

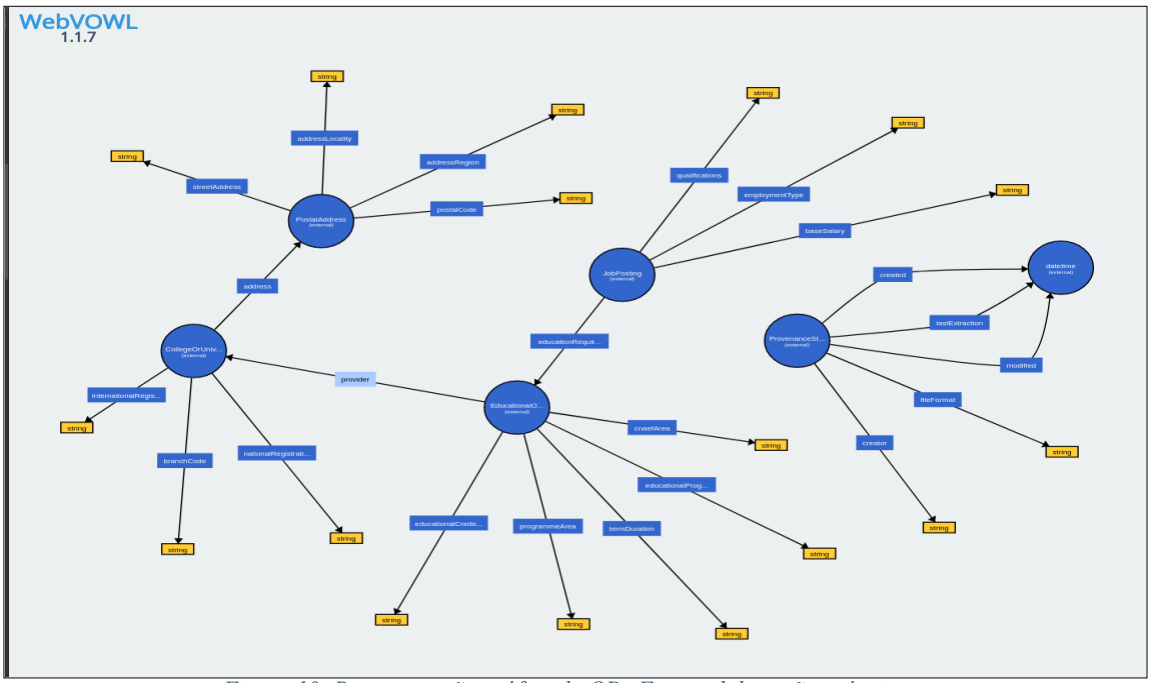

Figura 10: Representação gráfica da OP «Fonte: elaboração própria»

[51] https://pt.wikipedia.org/wiki/RDF/XML [Consult. em 23/08/2021].
[52] Gerada *online* no *website:* http://www.visualdataweb.de/webvowl/#opts=sidebar=0;#file=ontologia_psesp.owl

6 Desenvolvimento do PSESP

Este Capítulo descreve como foi feito o desenvolvimento do PSESP[53].

Na busca por *softwares* e ferramentas para desenvolver o PSESP, optou por utilizar os de código aberto[54], sendo assim, foram escolhidas como ferramentas principais: Python[55], como linguagem de desenvolvimento, Postgresql[56], como base de dados, Django[57], como *web framework*[58], e outras bibliotecas auxiliares.

Para facilitar o desenvolvimento e futuras expansões, a arquitetura do PSESP é composta por módulos (ver Figura 11):

1 Módulo de extração, transformação e consolidação, que entrega um conjunto de tabelas (TR's) populadas com informações extraídas da *web.*

2 Módulo de LDM para integração semântica, que liga as informações armazenadas nas TR e entrega a ontologia OP, tornando-a instanciada.

3 Módulo do Portal que disponibiliza os dados instanciados em formato aberto, que é uma página na *web* com os dados da ontologia OP, em formato aberto. Neste módulo também é disponibilizado uma página com interface *user friendly,* que permitirá ao utilizador executar uma consulta SPARQL diretamente na ontologia instanciada.

[53] https://github.com/LuSoMaBra/semantic-esp
[54] https://pt.wikipedia.org/wiki/Software_de_c%C3%B3digo_aberto [Consult. em 23/08/2021].
[55] https://www.python.org/ [Consult. em 23/08/2021].
[56] https://www.postgresql.org/ [Consult. em 23/08/2021].
[57] https://www.djangoproject.com/ [Consult. em 23/08/2021].
[58] https://pt.wikipedia.org/wiki/Framework_para_aplica%C3%A7%C3%B5es_web [Consult. em 23/08/2021].

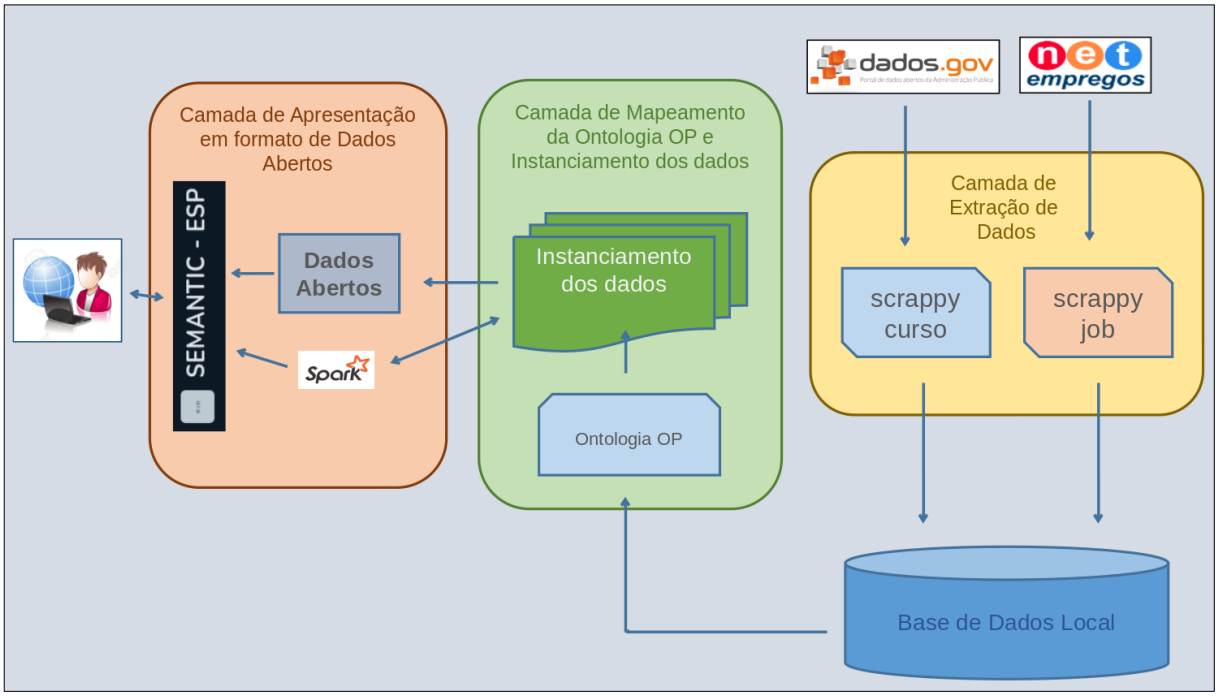

Figura 11: Arquitetura do PSESP «Idem»

A modularidade da estrutura permite que os componentes sejam substituídos à medida. Por exemplo, na camada de extração de dados poderia ser utilizada uma base de dados noutro formato, desde que contivesse as informações necessárias para satisfazer a ontologia.

A construção de um rastreador *web* que permite a recolha de dados, é responsável por extrair os dados de cursos e instituições e de saídas profissionais, tanto dos dados abertos governamentais (formato JSON) como de uma página da *web* (formato HTML) e é flexível ao ponto de poder ser usada para extrair dados noutros domínios apenas com adaptações no código. Além disso, o modelo de ontologia pode ser facilmente adaptado para uso em qualquer domínio.

No âmbito deste trabalho, não está prevista a criação de um controlador de execução dos módulos de extração, o que permitiria por exemplo, a automação desse processo. Somente a título de demonstração da funcionalidade, a execução dos módulos extratores foi implementada no portal de apresentação, porém esta não é uma boa prática se o PSESP ou derivação do mesmo, for implementado fora do âmbito académico.

As seções a seguir discutem cada módulo com mais detalhes.

6.1 Camada de extração de dados

A camada de extração de dados é responsável por recolher informação relevantes das fontes de informação e armazenar em base de dados própria com vista a consulta e tratamento posterior. Diferentes fontes de informação foram usadas para estruturação dessa camada. No presente trabalho apenas informações que estão disponíveis por meio de fontes públicas e de fácil acesso foram consideradas.

A próxima seção mostra em detalhes a criação de uma estrutura de base de dados para armazenar as informações extraídas da *internet*. De seguida, a seção 6.1.2 apresenta a extração dos dados de cursos e instituições. A seção 6.1.3 apresenta a extração de dados das ofertas de trabalho e por fim, seção 6.1.4 apresenta a integração dos dados extraídos bem como o armazenamento na base de dados.

6.1.1 Criação das Tabelas Relacionais

Para facilitar o processo de mapeamento, na definição da estrutura de entidades e relacionamentos da base de dados, optou-se por manter as tabelas e atributos em harmonia com a ontologia OP. Assim as classes da OP correspondem às tabelas, as propriedades de tipos de dados correspondem aos atributos e as propriedades dos objetos correspondem às chaves estrangeiras. Cada registo da tabela é unicamente identificado ou seja, tem como chave primária o atributo *id*, o qual é gerado automaticamente pela base de dados e será usado para a construção do URI que identificará cada instância da ontologia. A Figura 12 apresenta o esquema relacional desenvolvido, o qual foi criado com a ferramenta dbeaver[59].

[59] https://dbeaver.io/

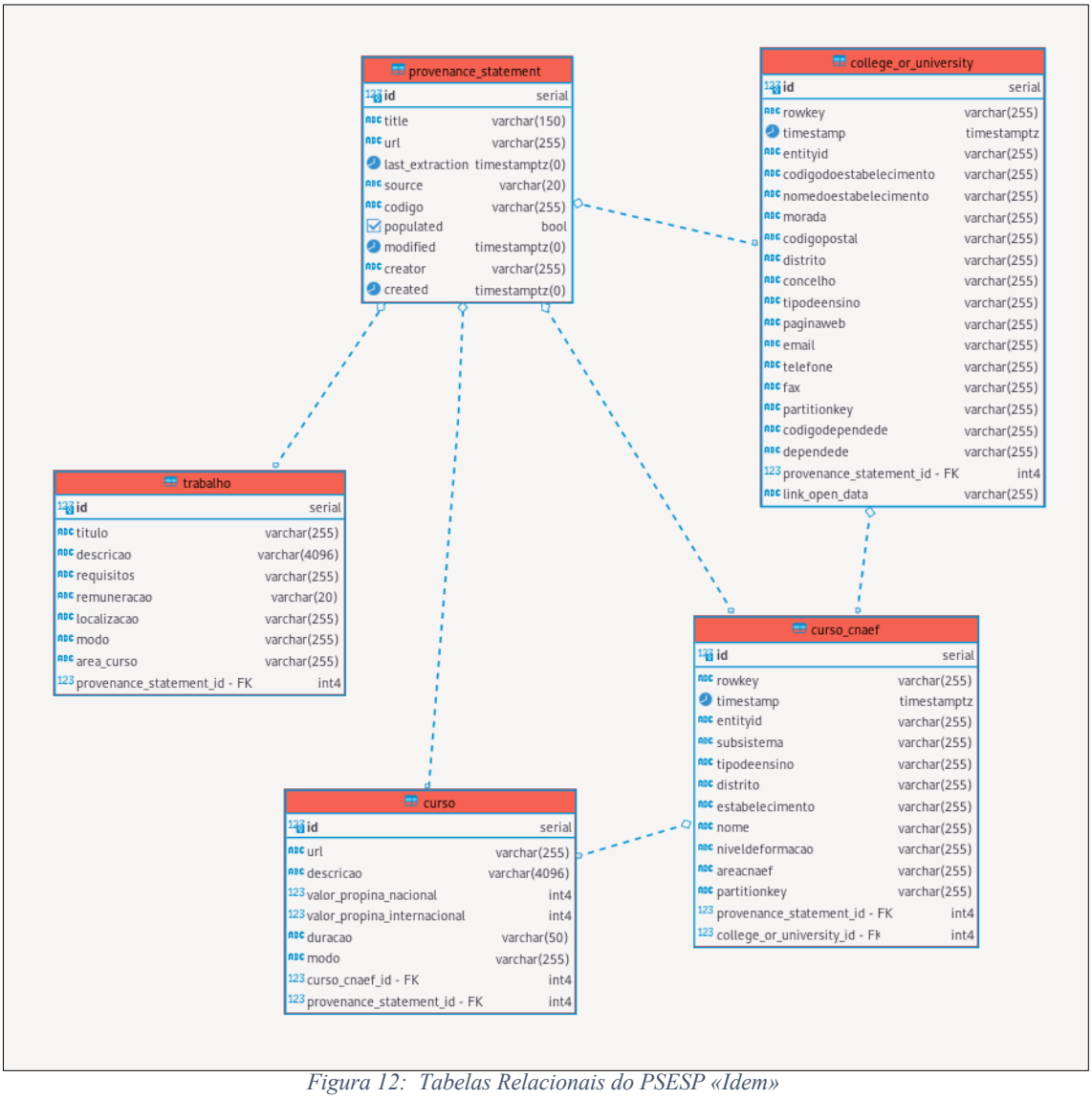

Figura 12: Tabelas Relacionais do PSESP «Idem»

Na Figura 12 são apresentadas cinco tabelas, as quais são descritas como segue:

- A tabela *provenance_statement* é preenchida com dados de metadados de informações das extrações e guarda a proveniência dos dados das outras tabelas. Um breve dicionário de dados desta tabela é apresentado na Tabela 6.
- A tabela *college_or_university* armazena as informações relacionadas às instituições. É preenchida com dados oriundos das bases de dados abertas do governo, conforme os endereços vistos no Capítulo 6 (seção 6.1.2). Nem todos os atributos da tabela são interessantes no âmbito deste trabalho, portanto, no breve dicionário de dados apresentado na tabela 7 são mostrados apenas os atributos utilizados.

- A tabela *curso_cnaef* armazena as informações relacionadas aos cursos de cada instituição. Da mesma forma, esta tabela é também preenchida com dados oriundos das bases de dados abertas do governo. Da mesma forma que ocorreu na tabela *curso_cnaef, n*em todos os atributos são interessantes no âmbito deste trabalho, portanto, no breve dicionário de dados apresentado na tabela 8 são mostrados apenas os atributos utilizados.
- A tabela *curso* é preenchida com dados recolhidos pela extração de cursos. Esta tabela guarda informações adicionais importantes no âmbito deste trabalho que não foram encontradas nos dados abertos do governo. Um breve dicionário de dados desta tabela é apresentado na tabela 9. Os atributos desta tabela poderiam estar incluídos na tabela *curso_cnaef*, porém foi criada uma tabela separada para armazenar os dados devido a volatibilidade da natureza dos mesmos.
- A tabela trabalho é preenchida com informações extraídas do *website* das ofertas de trabalho. Um breve dicionário de dados desta tabela é apresentado na Tabela 10.

O código fonte na sintaxe PostgreSQL[60] está no Anexo 01 – Código fonte das TR's.

Tabela 6 – Dicionário de dados da tabela provenance_statement «Idem»

Atributos	Descrição
title	Nome dado aos dados da extração. Ex.: "Instituições de Ensino Superior", "Classificação Nacional", "Extrator de dados da Universidade Trás-os-Montes", "Extrator de dados do website net-empregos.com"
url	Endereço eletrônico de extração. Ex.: "https://dados.gov.pt/pt/datasets/r/59ed02b9-410c-4f68-81ef-a3755ca66400"
last_extration	Data da última extração.
source	Tipo dos dados extraídos. Ex.: "JSON", "HTML"
codigo	Código interno. Ex: "InstituicoesdoEnsinoSuperior", "UTAD_Spider", etc.
populated	Indica se foi extraído ou não. Ex.: True/False
creator	Nome da entidade que disponibiliza os dados. Ex.: "Portal de dados abertos da Administração Pública (https://dados.gov.pt)", etc.
created	Data da criação dos dados.
Modified	Data da alteração dos dados.

[60] https://pt.wikipedia.org/wiki/SQL [Consult. em 23/08/2021].

Tabela 7 – Dicionário de dados da tabela college_or_university «Idem»

Atributos	Descrição
codigodoestabelecimento	Código da instituição.
nomedoestabelecimento	Nome da instituição. Ex.: Academia da Força Aérea
morada	Morada da instituição.
codigopostal	Código postal da instituição. Ex.: 2563-251
distrito	Distrito da instituição.
concelho	Concelho da instituição. Ex.: Lisboa, Coimbra
provenance_statement_id	Chave estrangeira para a tabela *provenance_statement.*

Tabela 8 – Dicionário de dados da tabela curso_cnaef «Idem»

Atributos	Descrição
estabelecimento	Código da instituição.
nome	Nome do curso. Ex.: 5662 - Ciências da Linguagem
niveldeformacao	Grau associado ao curso. Ex.: Licenciatura 1º Ciclo, Mestrado
area_cnaef	Área do curso. Ex.: 640 - Ciências Veterinárias
college_or_university_id	Chave estrangeira para a tabela *college_or_university.*
provenance_statement_id	Chave estrangeira para a tabela *provenance_statement.*

Tabela 9 – Dicionário de dados da tabela curso «Idem»

Atributos	Descrição
url	*Link* para acesso direto ao Curso.
descricao	Descrição do Curso.
valor_propina_nacional	Valor do curso anual por aluno nacional.
valor_propina_internacional	Valor do curso anual por aluno internacional.
duracao	Duração do curso em semestres.
modo	Modo de ensino. Ex.: Tempo Integral, noturno, etc.
curso_cnaef_id	Chave estrangeira para a tabela *curso_cnaef*
provenance_statement_id	Chave estrangeira para a tabela *provenance_statement.*

Tabela 10 – Dicionário de dados da tabela trabalho «Idem»

Atributos	Descrição
titulo	Título da oferta de emprego. Ex.: Engenheiro de Informática, Gerente de Micro Serviços, etc.
descricao	Descrição do trabalho.
localizacao	Localização onde será efetuado o trabalho.
area_curso	Área de curso que serve de ligação ao curso. Ex.: Engenharia Informática, Gestão e Marketing, Turismo, etc.
remuneracao	Valor do salário anual.
requisitos	Conhecimentos e habilidades necessárias para ocupação da vaga.
modo	Modo de trabalho oferecido. Ex.: Presencial, Remoto, *Full time*, etc.
provenance_statement_id	Chave estrangeira para a tabela *provenance_statement.*

6.1.2 Extração de dados de instituições e cursos

A extração de dados de instituições e cursos é feita em dois processos que diferem entre si pelas utilização das ferramentas utilizadas no código do PSESP:

1) Extração de dados abertos de instituições e cursos em formato JSON cujas bibliotecas são nativas no Python;

2) Extração dos dados adicionais dos cursos que estão disponíveis nos *website*s das instituições de ensino, ou seja, no formato HTML, utilizando a biblioteca "Scrappy"[61].

As informações sobre as instituições foram obtidas a partir do *website* de dados abertos da Administração Pública[62] no formato JSON. No documento disponível para descarga foram extraídas respetivamente informações de todas as instituições em território português registadas. Essas informações foram armazenadas na base de dados local, respeitando a mesma estrutura das chaves e valores do formato original do ficheiro JSON para os atributos das tabelas. Por exemplo, o valor da chave *rowkey* é armazenado no atributo *rowkey* da tabela, o valor da chave *entityid* é armazenado no atributo *entityid* da tabela equivalente e assim por diante com todas as chaves e atributos. Optou-se por armazenar todos os dados do ficheiro JSON, a despeito de, no âmbito deste trabalho, não serem todos utilizados.

[61] https://scrapy.org/ [Consult. em 23/08/2021].
[62] https://dados.gov.pt/pt/ [Consult. Em 23/08/2021].

As informações sobre os cursos foram obtidas também a partir do *website* de dados abertos da Administração Pública[63] no formato JSON. No documento disponível para descarga foram extraídas respetivamente informações de todas os cursos de todas instituições em território português. Essas informações foram armazenadas na base de dados local, respeitando a mesma estrutura das chaves e valores do formato original do ficheiro JSON para os atributos das tabelas. Por exemplo, o valor da chave *rowkey* é armazenado no atributo *rowkey* da tabela equivalente, o valor da chave *partitionkey* é armazenado no atributo *partitionkey* da tabela e assim por diante com todas as chaves e atributos. Da mesma forma, optou-se por armazenar todos os dados do ficheiro JSON, mesmo que, no âmbito deste trabalho, não sejam todos utilizados.

As informações que foram armazenadas na tabela curso não estão disponíveis nos dados abertos do governo e, portanto, foram extraídas dos *websites* de cada curso a partir do formato HTML usando a biblioteca "Scrappy". O atributo *curso_cnaef_id* é uma chave estrangeira (PK) para o atributo *id* da tabela *curso_cnaef*, assegurando assim, a relação unívoca entre as duas tabelas.

Como a recolha das informações complementares oriundas dos *websites* das instituições de ensino não é uniforme, visto cada instituição disponibilizar e estruturar a informação de seus cursos conforme achar mais adequado, é necessário uma customização do código para extração de dados de cada universidade ou instituição. Para o presente trabalho, optou-se por construir somente o extrator da instituição "Universidade Trás-os-Montes e Alto Douro"[64], a qual serve como exemplo para a construção de extratores de dados de outras instituições.

O código do módulo de extração de cursos está dentro do ficheiro "scrappy_curso.py". Para adicionar os extratores de outras instituições, basta inserir dentro desse ficheiro os códigos específicos para cada uma, semelhante ao criado para a "Universidade Trás-os-Montes e Alto Douro" (UTAD).

A execução do módulo extrator do curso é feita manualmente através da linha de comando: "scrapy runspider scrapies-esp/scrapy_curso.py". Entretanto, apenas para efeitos de demonstração académica, a chamada ao módulo foi incluída diretamente no portal PSESP. Em extensões deste trabalho é altamente recomendável que seja criado um controlador de execução de módulos de extração.

---

[63] https://dados.gov.pt/pt/ [Consult. em 23/08/2021].
[64] https://www.utad.pt/estudar/inicio/licenciaturas-mestrados-integrados/ [Consult. em 23/08/2021].

Para detalhes mais técnicos, o código fonte na linguagem Python pode ser visto no Anexo 02 – Código fonte do extrator da UTAD.

6.1.3 Extração de dados de saídas profissionais

As informações sobre as saídas profissionais (ofertas de trabalho) foram extraídas do *website* "net-empregos.com". Estes dados estão no formato HTML, portanto, o extrator de dados de saídas profissionais considerou os mesmos princípios utilizados na extração das informações adicionais dos cursos, utilizando também a biblioteca "Scrappy".

O rastreador *web* recolhe os dados na página do "net-empregos.com" e processa os dados recolhidos, fazendo a separação das informações relevantes para só então guardá-las na base de dados. Nesse ponto é interessante mencionar que, ao recolher as informações sobre as saídas profissionais, o rastreador *web* filtra a sua pesquisa a partir das informações de cursos armazenadas na base de dados oriundas do processo de extração de cursos. Desta forma, a integração ocorre antes do instanciamento propriamente dito, trazendo resultados mais assertivos. A integração será apresentada em detalhe no Capítulo 6 (seção 6.1.4).

O *link* base para extração de dados de saídas profissionais é "https://www.net-empregos.com/pesquisa-empregos.asp?"[65], entretanto, devido a diversidade de ofertas de emprego que abrangem áreas que não interessam ao âmbito deste trabalho, foi necessária a filtragem com alguns elementos, a saber, "Formação Superior", para garantir que as ofertas correspondam a cursos superiores e "Tempo Integral¨, para excluir ofertas *part time* e temporárias.

O código do módulo de extração de ofertas de trabalho está dentro do ficheiro "scrapy_job_net_empregos.py". Para adicionar os extratores de outros *websites* de ofertas de trabalho, basta inserir dentro desse ficheiro os códigos específicos para cada uma, semelhante ao criado para "net-empregos.com". Quando for executado, será feita a extração de todas as informações de cada um automaticamente.

Semelhante ao mencionado na seção 6.1.3, a execução do módulo extrator de ofertas de trabalho é feita manualmente através da linha de comando, alterando apenas o parâmetro do

[65] https://www.net-empregos.com/pesquisa-empregos.asp?chaves=Forma%E7%E3o+Superior&cidade=&categoria={}&zona=0&tipo=1 [Consult. em 23/08/2021].

ficheiro: “scrapy runspider scrapies-esp/scrapy_job_net_empregos.py”. Para efeitos de demonstração académica, a chamada ao módulo foi incluída diretamente no portal PSESP.

6.1.4 Integração dos dados

No âmbito deste trabalho faz-se necessária a integração dos dados outrora extraídos separadamente. Esta integração objetiva dar valor às informações extraídas de modo a permitir atingir os objetivos propostos. Para obter como resultado os dados dos cursos integrados às saídas profissionais, optou-se por extrair as informações de ofertas de trabalho a partir das informações de cursos de maneira que o armazenamento dos dados foi feito considerando a integração dos mesmos.

Foi construído um algoritmo no código do PSESP que torna o processo dinâmico. Para funcionar, os dados do atributo *areacnaef*, da tabela *curso_cnaef* é armazenado na tabela *trabalho* no atributo *area_curso*. Desta forma, mesmo não sendo chave estrangeira definida na estrutura da base de dados, será possível extrair dados integrados das duas tabelas quando for necessário.

O módulo extrator para a tabela *trabalho* pesquisa na base de dados todos os dados não repetidos do atributo *area_curso* de todos cursos extraídos e aplica cada resultado ao filtro na extração de saídas profissionais. Com isto, esse módulo cria dinamicamente um *link* de extração de ofertas da referida área, executa a extração de todos os *links* apresentados no resultado da pesquisa e guarda na base de dados.

Na extração das ofertas de emprego, foi identificado que o *website* “net-empregos.com” trabalha com uma estrutura própria de áreas de emprego, um fragmento deste ficheiro, em fomato JSON[66], pode ser visto na Figura 13.

[66] https://pt.wikipedia.org/wiki/JSON [Consult. em 23/08/2021].

{ '29': 'Administração / Secretariado', '39': 'Agricultura / Florestas / Pescas', '22': 'Arquitectura / Design', '40': 'Artes / Entretenimento / Media', '16': 'Banca / Seguros / Serviços "Financeiros', '47': 'Beleza / Moda / Bem Estar', '57': 'Call Center / Help Desk', '53': 'Comercial / Vendas', '8': 'Comunicação Social / Media', '51': 'Conservação / Manutenção / Técnica', '23': 'Construção Civil', '15': 'Contabilidade / Finanças', '28': 'Desporto / Ginásios', '44': 'Direito / Justiça', '11': 'Educação / Formação', '54': 'Engenharia (Ambiente)', '45': 'Engenharia (Civil)', '46': 'Engenharia (Eletrotecnica)', '24': 'Engenharia (Mecanica)', '50': 'Engenharia (Química / Biologia)', '41': 'Farmácia / Biotecnologia', '26': 'Gestão de Empresas / Economia', '32': 'Gestão RH', '9': 'Hotelaria / Turismo', '12': 'Imobiliário', '6': 'Indústria / Produção', '38': 'Informática (Analise de Sistemas)', '34': 'Informática (Formação)', '37': 'Informática (Gestão de Redes)', '35': 'Informática (Internet)', '36': 'Informática (Multimedia)', '5': 'Informática (Programação)', '49': 'Informática (Técnico de Hardware)', '56': 'Informática (Comercial/Gestor de Conta)', '58': 'Limpezas / Domésticas', '30': 'Lojas / Comércio / Balcão', '19': 'Publicidade / Marketing', '18': 'Relações Públicas', '42': 'Restauração / Bares / Pastelarias', '14': 'Saúde / Medicina / Enfermagem', '55': 'Serviços Sociais', '52': 'Serviços Técnicos', '1': 'Telecomunicações', '43': 'Transportes / Logística' }

Figura 13: Fragmento do ficheiro de categorias do "net-empregos.com" «Idem»

É importante frisar que a estrutura apresentada na Figura 13 é bem diferente da forma como as áreas de emprego são armazenadas na base de dados do presente trabalho. Para resolver este problema e permitir a pesquisa e associação de cursos às ofertas de emprego, foi criado um algoritmo simples que utiliza as palavras do atributo *area_curso* na pesquisa dessas palavras no conteúdo do ficheiro de categorias do *website* "net-empregos.com". Por exemplo, para a associação das ofertas de trabalho do curso de "Engenharia de Informática", o módulo procura individualmente as palavras "engenharia" e "informática" no ficheiro de categorias apresentado na Figura 13, e retorna os códigos de categorias equivalentes. Neste exemplo, o retorno seria a lista: ['38', '34', '37', '35', '36', '5', '49', '56']. A partir daí, o extrator cria dinamicamente para cada item da lista um *link* para pesquisa e efetua uma iteração em todos resultados encontrados por essa pesquisa e armazena os dados de cada oferta de trabalho na tabela *trabalho*.

A integração se concretiza quando é gravado valor do atributo *areacnaef* originário da tabela *curso_cnaef* no atributo *area_curso* da tabela *trabalho*, mantendo a integridade da base de dados na relação entre as referidas tabelas e permitindo o posterior instanciamento da ontologia OP.

A Figura 14 apresenta um fluxograma que descreve o algoritmo de extração dos dados de saídas profissionais incluindo a respetiva integração. Para detalhes mais técnicos, o código fonte na linguagem Python pode ser visto no Anexo 03 – Código fonte do extrator do *website* "net-empregos.com".

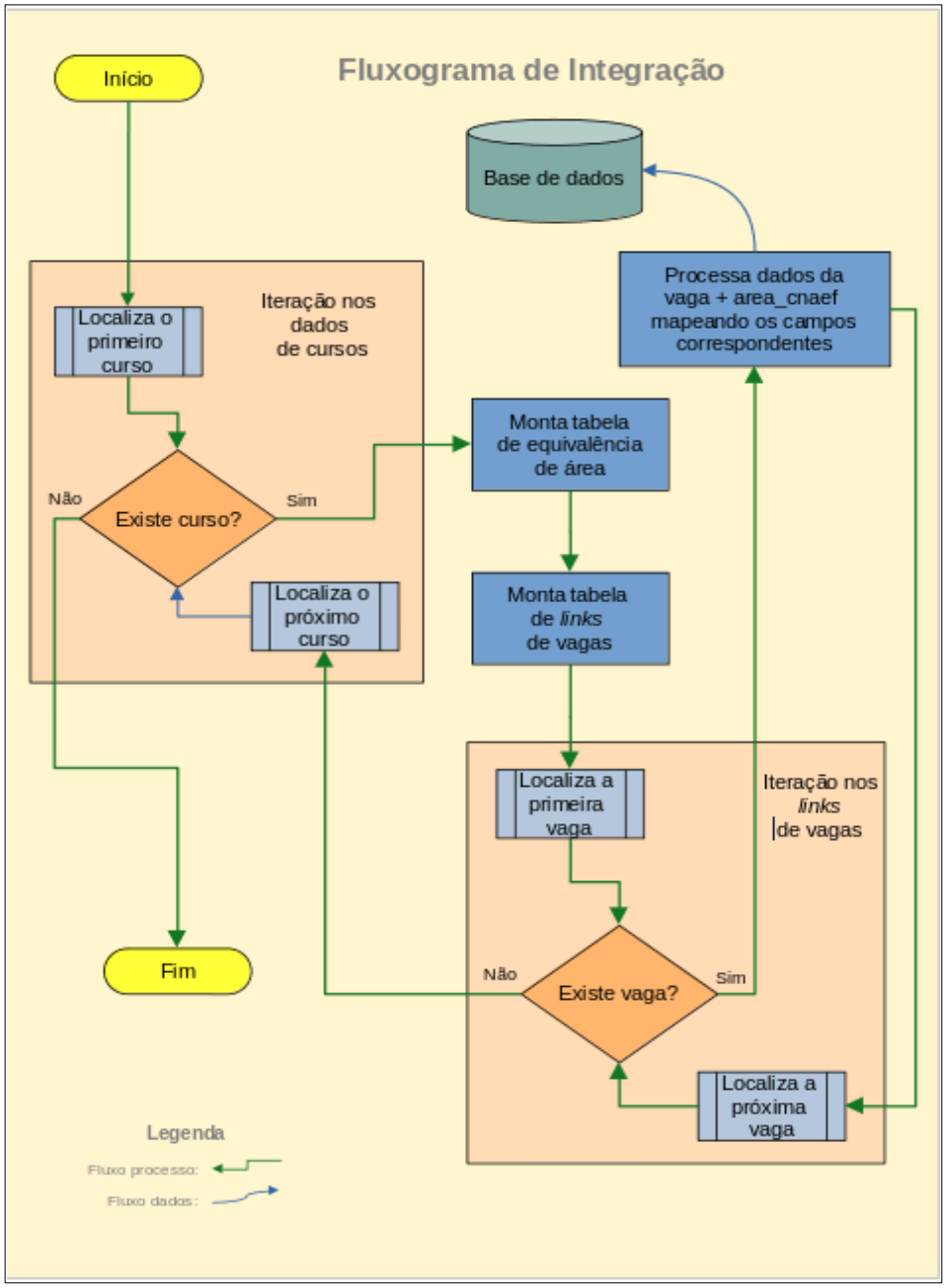

Figura 14: Fluxograma de integração «Idem»

6.2 Camada de mapeamento RDB para RDF

Essa seção descreve como foi feito o mapeamento RDB (TR) para RDF (ontologia OP). Nesta camada de mapeamento, o processo foi dividido em duas fases: 1) definição dos

mapeamentos entre a base de dados e a ontologia OP; e 2) execução dos mapeamentos de modo a obter os triplos RDF. Cada uma dessas fases é detalhada a seguir.

6.2.1 Definição dos mapeamentos entre RDB e RDF

Como já foi mencionado neste trabalho e brevemente explicado no Capítulo 4 (seção 4.3), optou-se por fazer um mapeamento direto entre as tabelas relacionais (TR) e a ontologia OP. Quer isso dizer que cada classe da ontologia corresponde a uma tabela da BD relacional e cada propriedade da ontologia corresponde ao atributo respetiva da tabela. Além disso, para facilitar o processo de mapeamento, os nomes dados às tabelas e respetivos atributos são mostrados na Tabela 11. Na Tabela 11 a coluna "Objetos" apresenta os objetos (tabelas ou atributos) da base de dados, a coluna "Termos" apresenta os termos da ontologia OP, e a coluna "Tipo" apresenta o tipo de relação entre os objetos e os termos.

É importante informar que, por não se estar a usar uma linguagem de mapeamento, apenas se consegue especificar o tipo de mapeamento envolvido e quem corresponde a quem. Entretanto, em alguns mapeamentos de propriedade serão necessários fazer transformações nos dados antes de serem instanciados na ontologia. Isto é feito a nível de código e brevemente será explicado na seção a seguir.

Tabela 11 – Tabela de Mapeamento RDB para ontologia OP «Idem»

Linha	Objetos	Termos	Tipo
1	curso e curso_cnaef	schema:EducationalOcuppationalProgram	tabela-classe
2	college_or_university	schema:CollegeOrUniversity	tabela-classe
3	provenance_statement	schema:ProvenanceStatement	tabela-classe
4	trabalho	schema:JobPosting	tabela-classe
5	Nome (tabela curso_cnaef)	schema:name	propriedade-propriedade
6	nomedoestabelecimento	schema:name	propriedade-propriedade
7	descricao (tabela trabalho)	schema:description	propriedade-propriedade
8	descricao (tabela curso)	schema:description	propriedade-propriedade
9	modo	schema:educationalProgramMode	propriedade-propriedade
10	title (tabela provenance_statement)	schema:title	propriedade-propriedade
11	title (tabela trabalho)	schema:title	propriedade-propriedade
12	codigodoestabelecimento	schema:branchCode	propriedade-propriedade
13	url (tabela curso)	schema:url	propriedade-propriedade
14	url (tabela provenance_statement)	schema:url	propriedade-propriedade
15	morada	schema:streetAddress	propriedade-propriedade
16	concelho	schema:addressLocality	propriedade-propriedade
17	distrito	schema:addressRegion	propriedade-propriedade
18	codigopostal	schema:postalCode	propriedade-propriedade
19	creator	dcterms:creator	propriedade-propriedade
20	created	dcterms:createdpropriedade-propriedade	propriedade-propriedade
21	modified	dcterms:modified	propriedade-propriedade
22	duracao	schema:termDuration	propriedade-propriedade
23	niveldeformacao	schema:educationalCredentialAwarded	propriedade-propriedade
24	requisitos	schema:qualifications	propriedade-propriedade
25	remuneracao	schema:baseSalary	propriedade-propriedade
26	modo	schema:employmentType	propriedade-propriedade
27	source	schema:fileFormat	propriedade-propriedade
28	areacnaef	pseps:programmeArea	propriedade-propriedade
29	valor_propina_internacional	pseps:internationalRegistrationFee	propriedade-propriedade
30	valor_propina_nacional	pseps:nationalRegistrationFee	propriedade-propriedade
31	last_extraction	pseps:lastExtraction	propriedade-propriedade

6.2.2 Instanciamento da ontologia a partir da base de dados relacional

A segunda fase do processo de mapeamento é sua aplicação de modo a obter os triplos RDF. Esta tarefa ocorre quando é feito o instanciamento dos dados na ontologia OP no módulo

de mapeamento e instanciamento. Uma esquematização do funcionamento da camada de mapeamento pode ser vista na Figura 15.

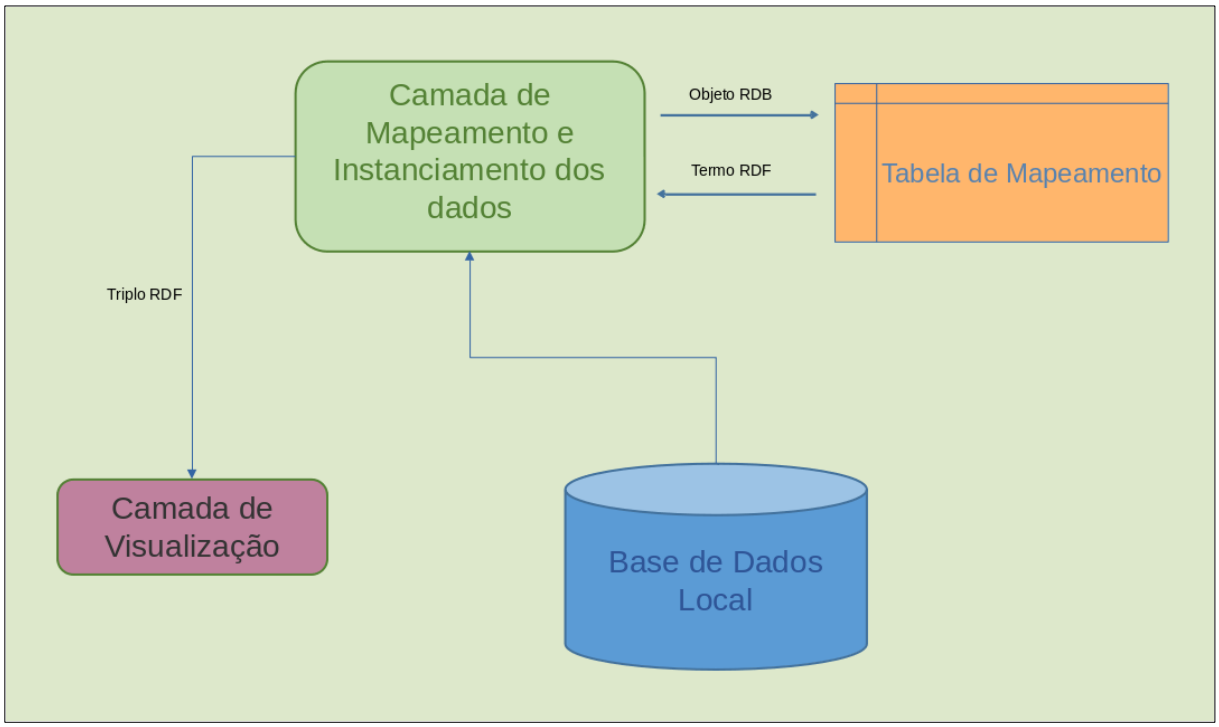

Figura 15: Esquema do mapeamento do PSESP «Idem»

Em síntese, a camada de mapeamento e instanciamento do PSESP lê os registos no RDB, consulta a tabela de mapeamento para localizar o termo, e faz o instanciamento do triplo. Em outras palavras, é feita uma iteração de leitura dos dados da base de dados local e a cada registo é consultada a tabela de mapeamento para identificar os termos equivalentes. Logo a seguir, é feito o instanciamento da ontologia com base nos dados RDB, fazendo as devidas transformações nesss dados quando necessário.

É importante ressaltar que a ontologia depois de instanciada, não é gravada em nenhum ficheiro, ou seja, a ontologia é instanciada em tempo real, quando o utilizador solicita, através da visualização.

Na camada de mapeamento, foi utilizada uma biblioteca do Python[67] chamada rdflib[68], que é um pacote para manipulação de RDF[69] e inclui recursos para serialização em vários

[67] https://pt.wikipedia.org/wiki/Python [Consult. em 23/08/2021].
[68] https://rdflib.readthedocs.io/en/stable/ [Consult. em 23/08/2021].
[69] https://en.wikipedia.org/wiki/Resource_Description_Framework [Consult. em 23/08/2021].

formatos, a saber, RDF/XML[70], NTriples[71], Turtle[72], JSON-LD[73], e outros, bem como a implementação de ontologias básicas do RDF.

Os triplos RDF criados pela camada de mapeamento e instanciamento são formados pelos sujeitos, que são os URI's (identificadores únicos dos registos de cada tabela), pelos predicados, que são as propriedades de dados mapeadas, e pelos objetos, que são os dados encontrados nos atributos das tabelas.

Foi designado um identificador único diferente para cada registo de cada tabela para evitar que na ontologia instanciada houvesse duplicidade de dados. Assim, o nome do URI é um texto composto pelo prefixo da ontologia, o nome da tabela, o texto "id" e o valor do atributo *id* de cada tabela.

Por exemplo, para a tabela *college_or_university*, considere os primeiros três registos cujos valores do atributo *id* são: 1, 2 e 3. Considere também que que o prefixo para a ontologia PSESP:https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp é *ontologia_psesp.* Assim o URI de cada um deles ficaria assim definido:

- ontologia_psesp/college_or_university_id_1
- ontologia_psesp/college_or_university_id_2
- ontologia_psesp/college_or_university_id_3

Uma vez entendido como o URI foi definido, é possível apresentar um exemplo de instanciamento das propriedades de dados e classe. Para isso, considere os dados disponíveis nas Figuras 16 e 17. A Figura 16 mostra a tabela *curso_cnaef* com o registo identificado pelo id=6777 e os dados do curso "9147 - Gestão" da universidade "Universidade de Trás-os-Montes e Alto Douro". A Figura 17 mostra um registo da tabela *curso* com o id=7, o qual contém informações complementares do curso 9147.

[70] https://pt.wikipedia.org/wiki/RDF/XML [Consult. em 23/08/2021].
[71] https://en.wikipedia.org/wiki/N-Triples [Consult. em 23/08/2021].
[72] https://en.wikipedia.org/wiki/Turtle_(syntax) [Consult. em 23/08/2021].
[73] https://en.wikipedia.org/wiki/JSON-LD [Consult. em 23/08/2021].

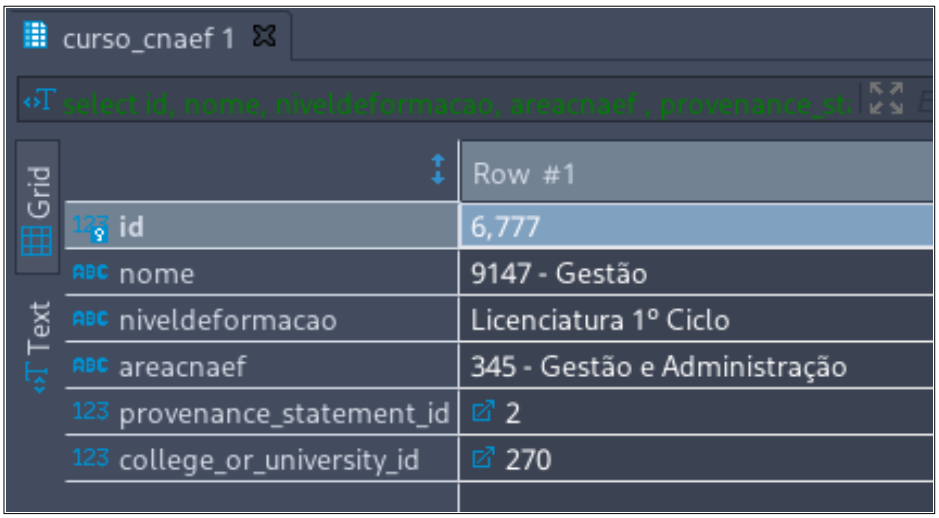

Figura 16: Exemplo de um registo da tabela curso_cnaef «Idem»

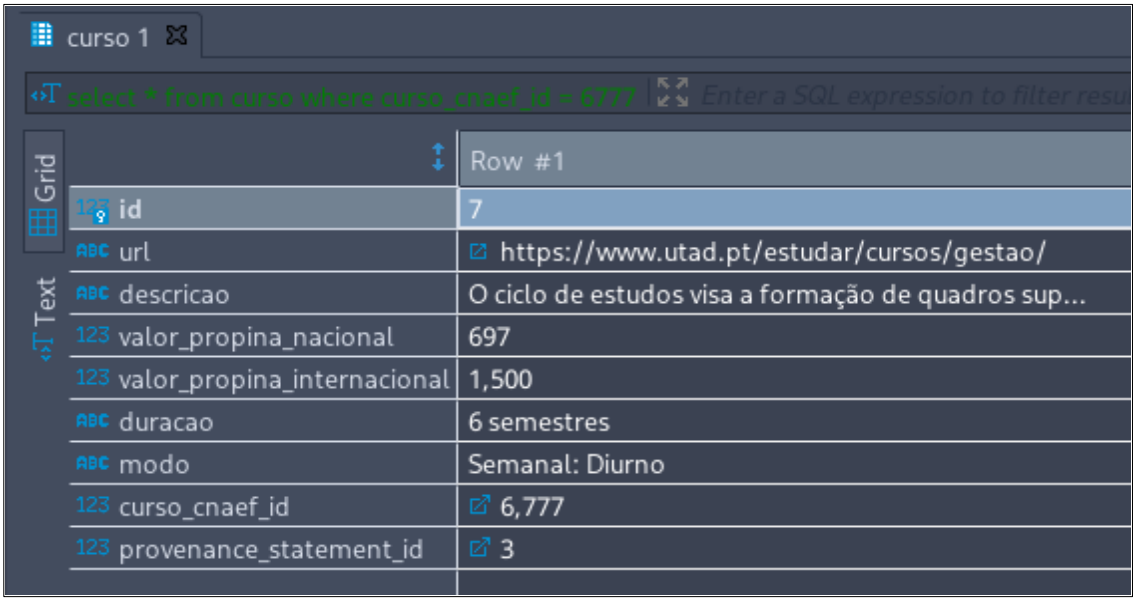

Figura 17: Exemplo de um registo da tabela curso «Idem»

Como mencionado na seção 6.2.2, a tabela *curso* possui como chave estrangeira o atributo *curso_cnaef_id*, que aponta para o atributo *id* da tabela *curso_cnaef,* formando uma única instância de registo de dados. Desta forma, ao aplicar o mapeamento do objeto *curso/curso_cnaef* (veja a Tabela 11, linha 1), tem-se como resultado o seguinte triplo RDF:

<ontologia_psesp/curso_cnaef_id_6739, rdfs:type, schema:EducationalOcuppationalProgram>

O mapeamento acima é um dos mais simples presentes neste trabalho, o qual não envolve uma propriedade de tipos de dados que não requer qualquer transformação no valor original do atributo correspondente, entretanto, conforme mencionado na Seção 6.2.1, alguns valores dos atributos tiveram que ser transformados antes de serem utilizados no instanciamento dos correspondentes termos da ontologia.

Essa transformação basicamente consiste em remover do valor original dados que foram considerados irrelevantes. Isso é o que acontece por exemplo, com o mapeamento entre o objeto *nome* e o termo *schema:name* (veja Tabela 11, linha 5). Considere por exemplo o atributo nome da tabela *curso_cnaef* apresentado na Figura 16. Para a geração do correspondente triplo RDF, o qual será o valor para o termo *schema:name*, foi retirada a primeira parte da string, ficando apenas a palavra “Gestão”. O Triplo resultante fica como se segue:

<ontologia_psesp/curso_cnaef_id_6739, schema:name, “Gestão”>

O mesmo processo ocorre também com o mapeamento entre o objeto *curso_cnaef* e o termo *schema:name* (veja Tabela 11, linha 28). Considere por exemplo, o atributo *area_cnaef* da tabela *curso_cnaef* apresentado na Figura 16. Para a geração do correspondente triplo RDF, o qual será o valor para o termo *pseps:programmeArea*, foi retirada a primeira parte da string, ficando apenas a palavra “Gestão e Administração”. O Triplo resultante fica como se segue:

<ontologia_psesp/curso_cnaef_id_6739, psesp:programmeArea, “Gestão e Administração”>

A Figura 18 apresenta os triplos RDF obtidos a partir dos registos mostrados nas Figuras 16 e 17. Note-se que o valor do termo “schema:description” é mostrado truncado por conter muitos caracteres, o que dificultaria a visualização.

Prefixo: ontologia_psesp:https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp
<ontologia_psesp/curso_cnaef_id_6739, rdfs:type, schema:EducationalOcuppationalProgram>
<ontologia_psesp/curso_cnaef_id_6739, rdfs:type, owl.NamedIndividual>
<ontologia_psesp/curso_cnaef_id_6739, schema:name, “Gestão”>
<ontologia_psesp/curso_cnaef_id_6739, schema:description, “O ciclo de estudos visa a formação de quadros sup...”>
<ontologia_psesp/curso_cnaef_id_6739, schema:educationalProgramMode, “Semanal: Diurno”>
<ontologia_psesp/curso_cnaef_id_6739, schema:url, “https://www.utad.pt/estudar/cursos/gestao/”>
<ontologia_psesp/curso_cnaef_id_6739, schema:termDuration, “6 semestres”>
<ontologia_psesp/curso_cnaef_id_6739, schema:educationalCredentialAwarded, “Licenciatura 1º Ciclo”>
<ontologia_psesp/curso_cnaef_id_6739, psesp:programmeArea, “Gestão e Administração”>

Figura 18: Exemplo de triplo RDF do curso «Idem»

Outro exemplo que pode ser verificado é aplicando a mesma lógica do exemplo anterior. Para isso, considere os dados disponíveis nas Figura 19. A Figura 19 mostra a tabela *college_or_university* com o registo identificado pelo id=270.

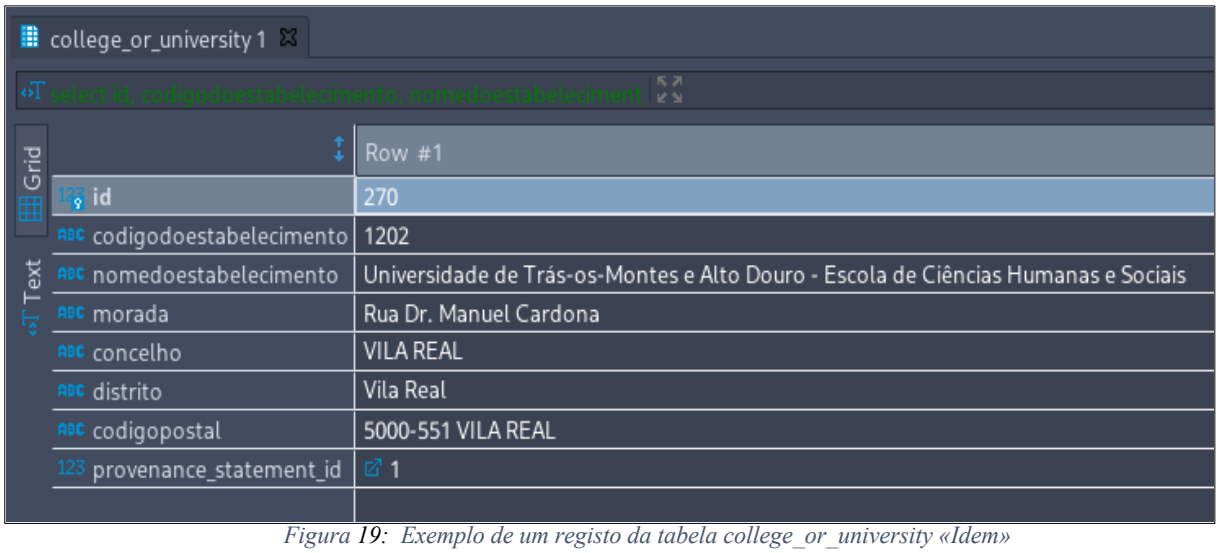

Figura 19: Exemplo de um registo da tabela college_or_university «Idem»

Desta forma, ao aplicar o mapeamento do objeto *college_or_university* (veja a Tabela 11, linha 2), tem-se como resultado o seguinte triplo RDF:

<ontologia_psesp/college_or_university_id_270, rdfs:type, schema:CollegeOrUniversity>

O mapeamento acima é tão simples quanto qualquer um que não envolve uma propriedade de tipos de dados ou requer transformação, no entanto, para propriedade de objeto do termo *schema:address,* cujo domínio é *schema:CollegeOrUniversity* (ver Tabela 3), é necessário instanciar os triplos relacionados. Assim, foi utilizado o texto identificado no objeto *codigopostal como parte da URI,* ficando conforme se segue:

<ontologia_psesp/schema:address_id_5000-551, rdfs:type, schema:PostalAddress>

O mesmo tipo de instanciamento é aplicado aos outros objetos do termo *schema:address* e por fim, para finalizar o exemplo, é feito o instanciamento de classe com o triplo formado pelo termo mapeado pela propriedade *schema:address* do *schema:CollegeOrUniversity,* tendo como resultado o triplo RDF:

<ontologia_psesp/college_or_university_id_270, schema:PostalAddress, ontologia_psesp/schema:address_id_5000-551>

A Figura 20 apresenta os triplos RDF obtidos a partir do registo mostrado na Figura 19.

Prefixo: ontologia_psesp:https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp

<ontologia_psesp/college_or_university_id_270, rdfs:type, schema:CollegeOrUniversity>
<ontologia_psesp/college_or_university_id_270, schema:branchCode, “1202”>
<ontologia_psesp/college_or_university_id_270, schema:name, “Universidade de Trás-os-Montes e Alto Douro - Escola de Ciências Humanas e Sociais”>
<ontologia_psesp/schema:address_id_5000-551, rdfs:type, schema:PostalAddress>
<ontologia_psesp/schema:address_id_5000-551, schema:postalCode, “5000-551 VILA REAL”>
<ontologia_psesp/schema:address_id_5000-551, schema:streetAddress, “Rua Dr. Manuel Cardona”>
<ontologia_psesp/schema:address_id_5000-551, schema:addressLocality, “VILA REAL”>
<ontologia_psesp/schema:address_id_5000-551, schema:addressRegion, “Vila Real”>
<ontologia_psesp/college_or_university_id_270, schema:PostalAddress, ontologia_psesp/schema:address_id_5000-551>

Figura 20: Exemplo de triplo RDF da instituição «Idem»

Com as informações extraídas, armazedadas, mapeadas e instanciadas, o passo seguinte é prover apresentação ao utilizador.

6.3 Camada de apresentação ao utilizador

Para a construção da camada de apresentação, foi utilizada a linguagem Python[74] juntamente com o *web framework* Django[75]. A escolha do Django deu-se pela facilidade do desenvolvimento de aplicações *web* com integração a qualquer base de dados, permitindo dinamismo na aplicação. Desta forma, foi possível o foco na escrita de código principal, com ganhos de tempo em funções já disponibilizadas no *framework*. No *front-end* foi utilizado *o template* bootstrap4[76] para prover um visual mais harmonioso e agradável na *User Interface* (UI).

O portal SESP consiste na apresentação de um ecrã principal de onde poderão ser acedidas as funcionalidades disponíveis ao utilizador. Cada funcionalidade abre um novo ecrã com as ações e informações pertinentes. Além disto, está disponível para todos ecrãs, um menu lateral em que podem ser acedidas as funcionalidades do ecrã principal sem necessariamente ter de voltar ao mesmo.

[74] https://www.python.org/ [Consult. em 23/08/2021].
[75] https://www.djangoproject.com/ [Consult. em 23/08/2021].
[76] https://www.w3schools.com/bootstrap/bootstrap_templates.asp [Consult. em 23/08/2021].

Como pode ser visto na Figura 14, há um texto que mostra os objetivos do portal, bem como botões de acesso às funcionalidades que são:

1. “Extração de Dados”, onde poderão ser processadas as extrações de dados de cursos ou de ofertas de trabalho, consoante a escolha do utilizador;

2. “Ontologias”, onde o utilizador terá opções de visualizar e/ou serializar e instanciar a ontologia OP;

3. “SPARQL”, onde o utilizador poderá criar e executar uma consulta padrão SPARQL sobre a ontologia instanciada.

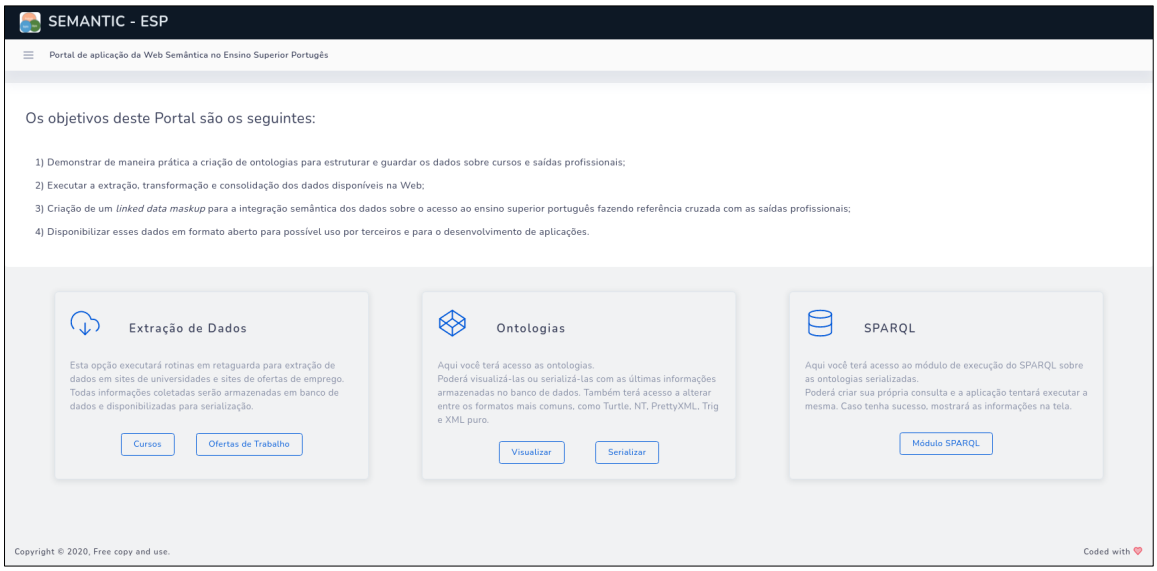

Figura 21: Ecrã principal do PSESP «Idem»

Cada ecrã de funcionalidades será explicado e detalhado a seguir.

6.3.1 Funcionalidade de Extração de Dados

Nesta funcionalidade, o portal irá executar a rotina de extração de dados.

Intuitivamente, cada escolha do utilizador neste ecrã irá chamar internamente as rotinas de extração explicadas no Capítulo 6 (seção 6.1). Como resultado, as TR’s da base de dados local serão preenchidas com as informações atuais recolhidas na extração.

A Figura 20 mostra o ecrã com uma lista de opções de extração. O utilizador poderá executar a extração equivalente ao carregar no botão “Extração” da linha escolhida. A opções são:

1. InstituicoesdoEnsinoSuperior: dados abertos de Instituições do Ensino Superior;

2. ClassNacionaldeareasdeeducacaoeformacao: dados abertos das áreas de educação e formação;
3. UTAD_spider: dados complementares da Universidade Trás-os-Montes e Alto Douro;
4. net-empregos: dados de saídas profissionais (ofertas de emprego) o *website* net-empregos.com.

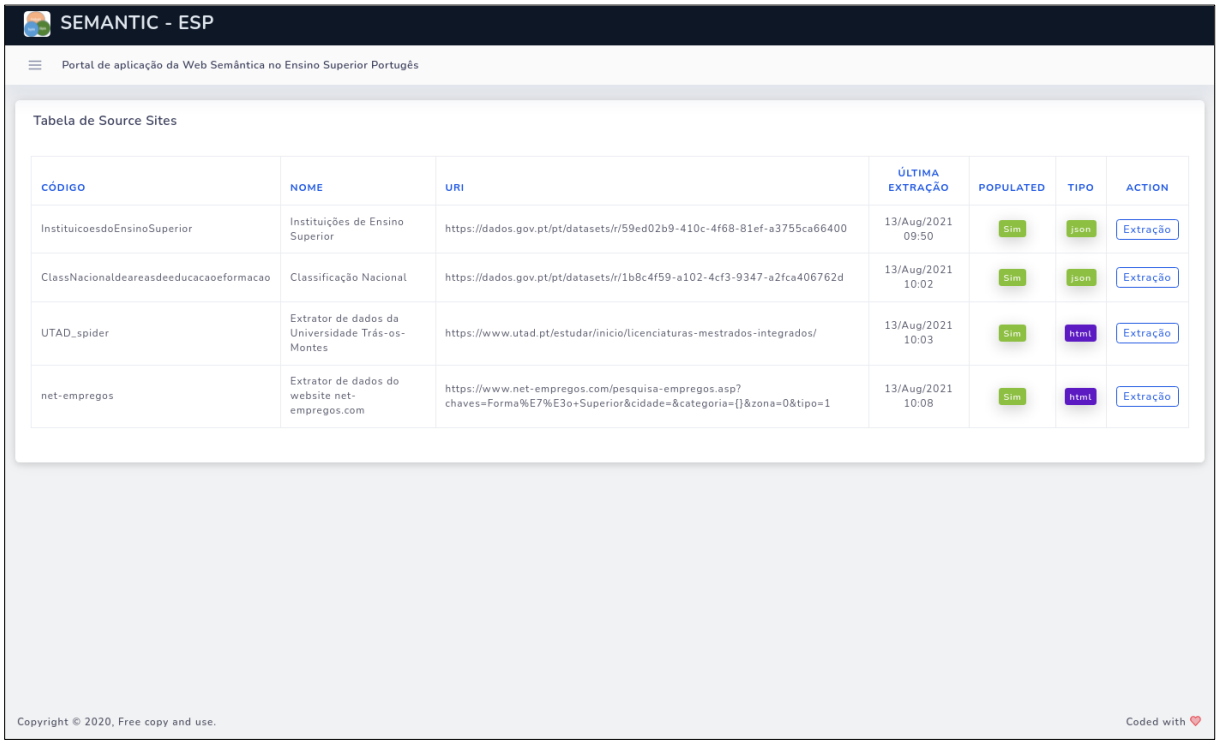

Figura 22: Ecrã de chamada de extração de dados do PSESP «Idem»

6.3.2 Funcionalidade de Ontologias

Nesta funcionalidade, ao carregar no botão “Visualizar” (Figura 14), o utilizador será remetido ao ecrã de visualização da ontologia OP conforme pode ser visto na Figura 16. Neste mesmo ecrã, poderá ser vista a ontologia OP no formato gráfico e no formato Turtle.

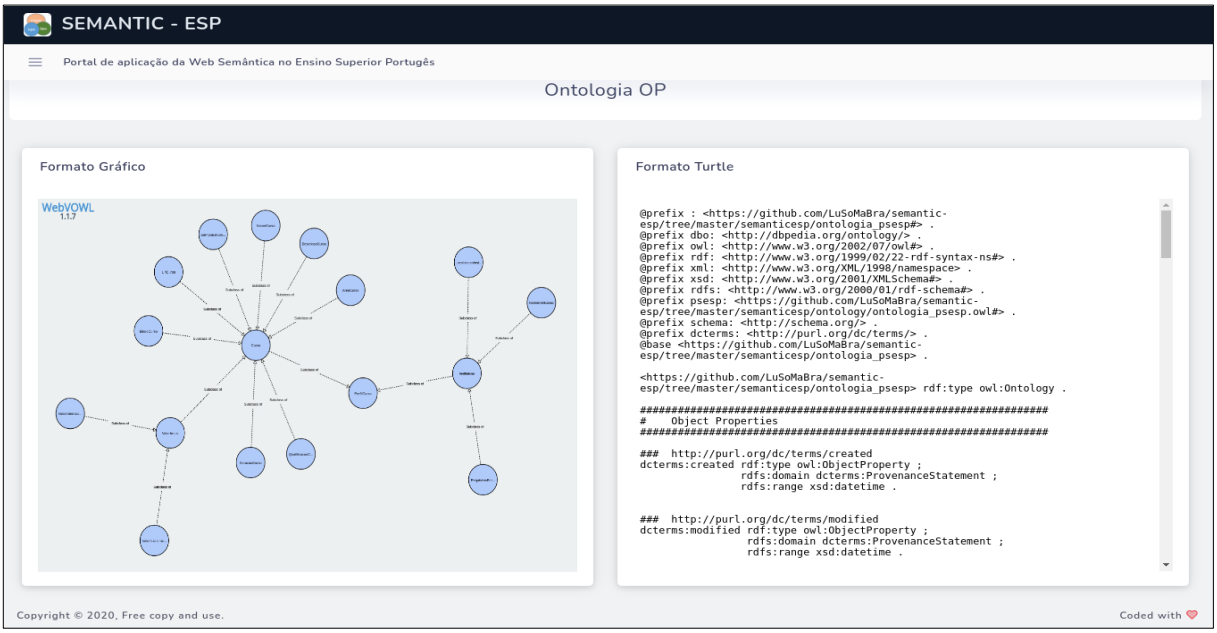

Figura 23: Ecrã de visualização da ontologia OP «Idem»

Ao carregar no botão “Serializar” (Figura 14), abre-se o ecrã de visualização da ontologia OP serializada e instanciada com as últimas informações armazenadas na base de dados. Neste mesmo ecrã, o utilizador tem a sua disposição as serializações mais comuns, tais como Turtle, NT, PrettyXML, Trig e XML, para visualizar as instâncias da ontologia criada (basta carregar no botão com o respetivo nome (veja Figura 24)). O utilizador tem ainda a funcionalidade de utilizar o *link* de acesso direto para descarregar as instâncias da ontologia criada. Assim, qualquer aplicação *web* que utilizar este *link* receberá de volta um ficheiro com o mesmo conteúdo das instâncias da ontologia visualizado no ecrã.

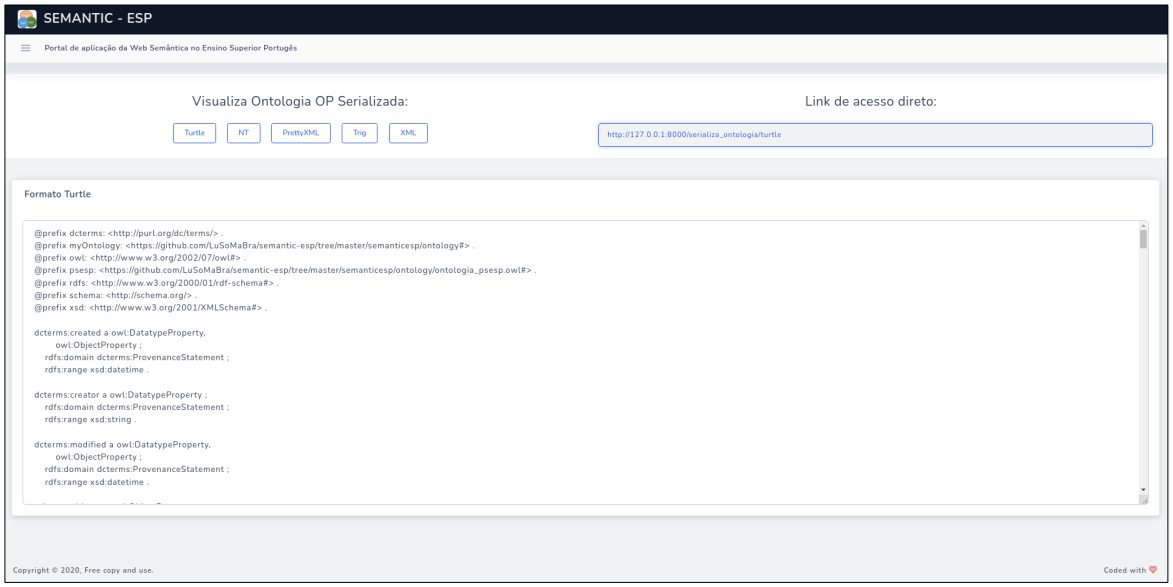

Figura 24: Ecrã de visualização da ontologia OP instanciada «Idem»

6.3.3 Funcionalidade do Módulo SPARQL

Nesta funcionalidade, o utilizador poderá criar sua própria consulta SPARQL e executar a sentença sobre a ontologia OP instanciada. Para isso, o utilizador escreve uma sentença apropriada no campo “Sentença SPARQL” e carrega no botão “Executar”. O resultado, se a sentença estiver correta, será apresentado imediatamente consoante os dados da ontologia OP.

A Figura 25 mostra um exemplo do resultado da sentença “SELECT ?x ?label WHERE { ?x owl:NamedIndividual ?label }” quando é processada através do botão “Executa”.

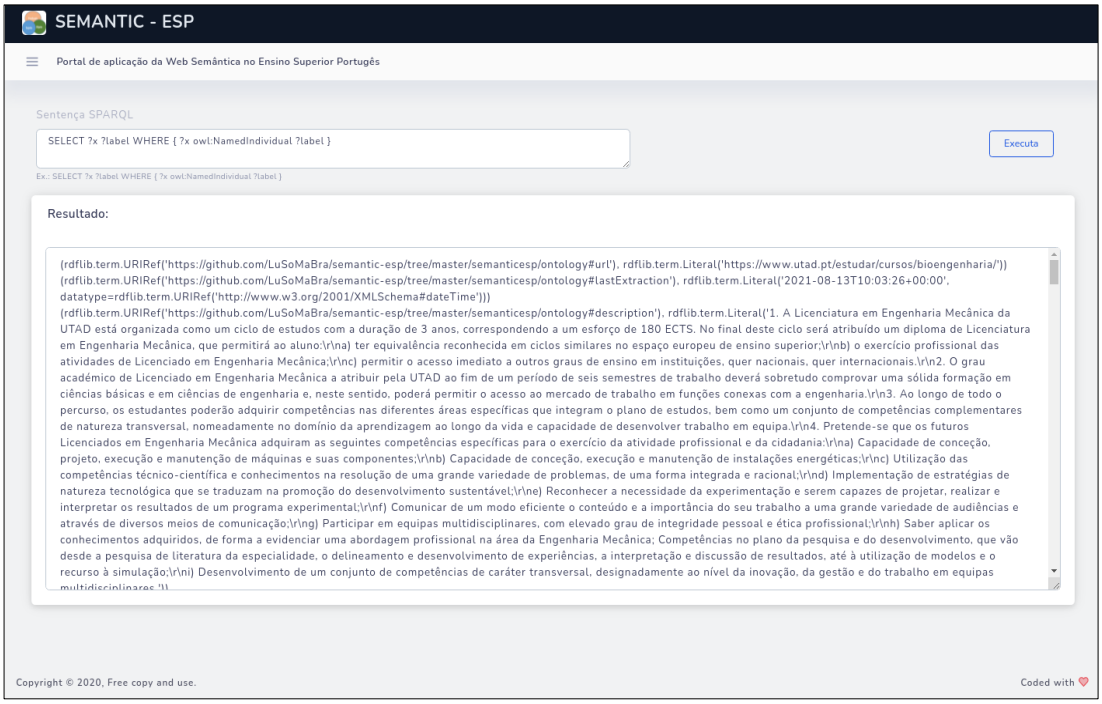

Figura 25: Ecrã de execução de consultas SPARQL «Idem»

7 Conclusão

A escolha da carreira profissional por parte de alunos prestes a cursarem algum curso superior e o enorme volume de dados sobre faculdades, cursos e ofertas de emprego, disponíveis na *internet,* fazem parte do cenário para o desenvolvimento deste trabalho.

Aparentemente o uso de tecnologias da w*eb semântica* como o LDM poderia ser a alternativa viável para superar estes problemas, o que de facto foi comprovado nos estudos avaliados na fundamentação teórica. A investigação do estado da arte trouxe o conhecimento de exemplos práticos desta tecnologia, o que permitiu definir como objetivo a criação de um LDM para integração semântica dos dados disponíveis na *web* referentes aos cursos oferecidos pelas instituições do ensino superior e os dados relativos às saídas profissionais, materializados na forma de ofertas de emprego disponíveis em *websites* especializados para o território português.

A proposta foi de criar um *linked data mashup* (LDM) para a integração de dados entre o acesso ao ensino superior português e as saídas profissionais previstas dos cursos disponíveis. Isso trouxe à tona algumas questões tais como: reunir as informações necessárias vindas de bases de dados externas ou mesmo *websites* e sem formato estruturado, manutenção da integridade e atualidade dos dados, e apresentação desses dados em formato aberto. Com isso, foi possível avançar na criação do PSESP, utilizando tecnologias da *web* semântica (LDM) e como resultado, foram obtidos em formatos diversos: do *website de* dados abertos do governo, dados sobre as universidades e cursos superiores, do *website* de uma universidade as universidades como exemplo, dados complementares dos cursos, e por fim, do *website* "net-empregos.com", dados de saídas profissionais, consoante aos cursos oferecidos pela universidade escolhida como exemplo. Os dados foram integrados semanticamente com junção das saídas profissionais o curso equivalente, e disponibilizados para uso segundo os padrões de dados abertos.

O portal criado (PSESP), disponibilizado na *web,* permite ver os resultados na prática. Para uso exclusivamente académico e voltado para utilizadores com conhecimentos de *web* semântica, o portal consegue executar a extração de dados a partir de diversas origens, a consolidação desses dados armazenamento em base de dados, o mapeamento de RDB para RDF, a integração dos dados de cursos e dados de saídas profissionais e por fim, a disponibilização em tempo real dos dados integrados no formato aberto.

Mesmo considerando que o foco do PSESP é o utilizador intermediário isto não deve ser considerada uma limitação e sim, que é possível estender para a criação de um portal voltado ao utilizador final, que poderá explorar, não só o resultado apresentado neste trabalho, mas também considerar outros fatores intrínsecos à escolha da carreira profissional tais como localização das universidades, bolsas de estudo, custo de vida enquanto estudante, oportunidades de emprego fora de Portugal, entre outros.

A principal contribuição foi a de abrir o caminho para utilização de dados em diferentes formatos e de fontes diversas, consolidá-las e integrá-las de forma a fornecer dinamicamente no formato da *web* semântica, conteúdos de que podem ser considerados informações de mais valia e não somente dados aleatórios.

8 Trabalhos Futuros

Neste Capítulo pretende-se propor sugestões para trabalhos futuros, levando em conta algumas questões e limitações encontradas no decorrer da criação deste trabalho.

Uma destas questões é saber se o facto de ter sido escolhida apenas uma universidade como exemplo de extração de dados, a despeito de todas outras existentes, poderia ter limitado o resultado final. Constatou-se que isso não se mostrou importante e não fez qualquer diferença significativa que impactasse o desenvolvimento deste trabalho.

Percebeu-se contudo que, com algum código adicional, poder-se-ia incluir várias outras instituições, o que permitiria enriquecer a base de dados para apresentação, porém não traria nenhum benefício imediato ao resultado esperado deste estudo estritamente académico. Entretanto, para além do mundo académico, numa extensão deste trabalho, seria necessária a inclusão de quantas mais instituições e cursos, incluindo os códigos de extração para cada uma delas no portal PSESP, de forma modular.

Esta, porém é uma abordagem frágil, uma vez que, caso a instituição resolva alterar a estrutura de seu *website*, o respetivo código de extração no PSESP, poderá ter que ser ajustado ou até mesmo ser reescrito, gerando algum prejuízo ou retrabalho no projeto.

Para resolver esta questão, deve ser estabelecido para trabalho futuro um processo de atualização periódico, que deveria considerar não só a atualização do código de cada módulo de universidade, caso houvesse alteração na fonte de dados, mas também as alterações dos dados extraídos.

Isso leva a outra questão importante. Atualmente, não há qualquer processo de manutenção de dados. Sempre que os dados são extraídos é criada uma nova versão do grado RDF sobrepondo-se ao existente. Isso, não é a solução ideal, o ideal seria fazer a manutenção incremental, ou seja, após cada extração, ver o que foi alterado e só atualizar o que foi alterado. Fazer a manutenção incremental está fora do escopo deste trabalho, e por ser algo complexo, por si só já seria tema para outro projeto.

Outra questão exposta no Capítulo 6 (seção 6.1) é relativa as chamadas aos módulos extratores. No âmbito deste trabalho, não está prevista a criação de um controlador de execução dos módulos de extração para a automação desse processo. Assim sendo, para extensão deste trabalho e numa eventual utilização fora do âmbito académico, é altamente recomendável que os módulos extratores sejam executados regularmente, num computador com as ferramentas

necessárias e as devidas ligações de rede. Isto poderá ser feito de forma automática, ou seja, no mesmo ambiente linux[77] em que são executados os serviços *web* e *database* do PSESP. Poderá também ser configurado como serviço a ferramenta "crontab"[78], que é um conjunto de comandos usados para executar tarefas regulares de agendamento nesse ambiente. Assim, os módulos de extração poderiam ser executados sem necessidade de intervenção manual do utilizador.

Outro assunto que se apresentou como dúvida no decorrer do desenvolvimento deste trabalho foi a definição de qual tipo de utilizador faria uso do portal PSESP. Optou-se por criar o portal com foco no utilizador com conhecimentos técnicos e não para o utilizador comum, a saber, o aluno do ensino pré-académico. Desta forma, foi possível dar foco nos processos de extração, mapeamento e integração dos dados visando atingir os objetivos propostos. De qualquer forma, o resultado deste trabalho permitirá aos utilizadores avançados a criação de aplicações específicas para os alunos que estão na altura de definir seu futuro profissional a partir de uma escolha académica. Como trabalhos futuros, considera-se que pode ser pertinente criar uma interface para os utilizadores finais, a saber, os alunos.

Ainda dentro desse mesmo assunto, fica como sugestão, a estensão deste trabalho materializada na criação de um portal voltado ao utilizador comum. Como agregação de valor, poderiam ser incluídas informações sobre médias de notas para entradas nas universidades, localização das faculdades de cada curso, custo médio de alimentação e hospedagem na região da instituição, bolsas e benefícios para alunos, além de outras informações relevantes que podem tornar o portal uma ferramenta de apoio, não só para o aluno, como também para as universidades e instituições portuguesas de ensino superior.

[77] https://www.linux.org/ [Consult. em 23/08/2021].
[78] https://www.guru99.com/crontab-in-linux-with-examples.html [Consult. Em 23/08/2021].

9 Bibliografia

[1] ALKHELIL, Abdul Himid. **The Relationship between Personality Traits and Career Choice: A Case Study of Secondary School Students**. International Journal of Academic Research in Progressive Education and Development, v. 5, n. 2, 16 maio 2016. Disponível em: https://doi.org/10.6007/ijarped/v5-i2/2132. Acesso em: 26 set. 2021.

[2] PARLAK, Adem; DEVELI, Sedat; EYI, Yusuf Emrah. **Factors affecting the choice of health specialty by medical graduates**. Medical Teacher, v. 37, n. 7, p. 702-703, 11 maio 2015. Disponível em: https://doi.org/10.3109/0142159x.2015.1042435. Acesso em: 26 set. 2021.

[3] KAMINSKY, Samuel E.; BEHREND, Tara S. **Career Choice and Calling**. Journal of Career Assessment, v. 23, n. 3, p. 383-398, 6 out. 2014. Disponível em: https://doi.org/10.1177/1069072714547167. Acesso em: 26 set. 2021.

[4] MOORJANI, J., MANIKA, M. S., & SUJATA, G. **Career choice and personality as predictors of cognitive interference**. In Journal of the Indian Academy of Applied Psychology, 3(2), 291-294. 2007.

[5] EDWARDS, K., & QUINTER, M. **Factors Influencing Students Career Choices among Secondary School students in Kisumu Municipality, Kenya**. In Journal of Emerging Trends in Educational Research and Policy Studies, 2 (2): 81-87. 2012.

[6] OGOWEWO, Bridget Oghenekome. **Factors Influencing Career Choice Among Secondary School Students: Implications for Career Guidance**. The International Journal of Interdisciplinary Social Sciences: Annual Review, v. 5, n. 2, p. 451-460, 2010. Disponível em: https://doi.org/10.18848/1833-1882/cgp/v05i02/59293. Acesso em: 27 set. 2021.

[7] MULLOLA, Sari *et al*. **Personality traits and career choices among physicians in Finland: employment sector, clinical patient contact, specialty and change of specialty**. BMC Medical Education, v. 18, n. 1, 27 mar. 2018. Disponível em: https://doi.org/10.1186/s12909-018-1155-9. Acesso em: 27 set. 2021.

[8] OGOWEWO, Bridget Oghenekome. **Factors Influencing Career Choice Among Secondary School Students: Implications for Career Guidance**. The International Journal of Interdisciplinary Social Sciences: Annual Review, v. 5, n. 2, p. 451-460, 2010.

Disponível em: https://doi.org/10.18848/1833-1882/cgp/v05i02/59293. Acesso em: 27 set. 2021.

[9] SINGARAVELU, Hemla D.; WHITE, Lyle J.; BRINGAZE, Tammy B. **Factors Influencing International Students' Career Choice.** Journal of Career Development, v. 32, n. 1, p. 46-59, set. 2005. Disponível em: https://doi.org/10.1177/0894845305277043. Acesso em: 27 set. 2021.

[10] COUPLAND, Christine. **Career definition and denial: A discourse analysis of graduate trainees' accounts of career**. Journal of Vocational Behavior, v. 64, n. 3, p. 515-532, jun. 2004. Disponível em: https://doi.org/10.1016/j.jvb.2003.12.013. Acesso em: 27 set. 2021.

[11] NEWTON, Dale A. **Trends in Career Choice by US Medical School Graduates**. JAMA, v. 290, n. 9, p. 1179, 3 set. 2003. Disponível em: https://doi.org/10.1001/jama.290.9.1179. Acesso em: 27 set. 2021.

[12] BURNS, Gary N. *et al*. **Personality, interests, and career indecision: a multidimensional perspective**. Journal of Applied Social Psychology, v. 43, n. 10, p. 2090-2099, 10 set. 2013. Disponível em: https://doi.org/10.1111/jasp.12162. Acesso em: 27 set. 2021.

[13] KIN, Lee Wai; RAMELI, Mohd Rustam Mohd. **Myers-Briggs Type Indicator (Mbti) Personality and Career Indecision among Malaysian Undergraduate Students of Different Academic Majors**. Universal Journal of Educational Research, v. 8, n. 5A, p. 40-45, maio 2020. Disponível em: https://doi.org/10.13189/ujer.2020.081906. Acesso em: 27 set. 2021.

[14] NOETH, RICHARD J.; ENGEN, HAROLD B.; NOETH, PATRICIA E. **Making Career Decisions: A Self-Report of Factors That Help High School Students**. Vocational Guidance Quarterly, v. 32, n. 4, p. 240-248, jun. 1984. Disponível em: https://doi.org/10.1002/j.2164-585x.1984.tb01587.x. Acesso em: 27 set. 2021.

[15] CROSNOE, Robert; CAVANAGH, Shannon; ELDER, Glen H. **Adolescent Friendships as Academic Resources: The Intersection of Friendship, Race, and School Disadvantage**. Sociological Perspectives, v. 46, n. 3, p. 331-352, set. 2003. Disponível em: https://doi.org/10.1525/sop.2003.46.3.331. Acesso em: 27 set. 2021.

[16] HANUSHEK, Eric *et al*. **Does Peer Ability Affect Student Achievement?**. Cambridge, MA: National Bureau of Economic Research, 2001. Disponível em: https://doi.org/10.3386/w8502. Acesso em: 27 set. 2021.

[17] ZIMMERMAN, David J. **Peer Effects in Academic Outcomes: Evidence from a Natural Experiment**. Review of Economics and Statistics, v. 85, n. 1, p. 9-23, fev. 2003. Disponível em: https://doi.org/10.1162/003465303762687677. Acesso em: 27 set. 2021.

[18] JACOBSEN, Mary H. **Hand-Me-Down Dreams: How Families Influence Our Career Paths and How We Can Reclaim Them**. [S. l.]: Three Rivers Press, 2000. 240 p. ISBN 9780609802649.

[19] HEARN, James C. **The Relative Roles of Academic, Ascribed, and Socioeconomic Characteristics in College Destinations**. Sociology of Education, v. 57, n. 1, p. 22, jan. 1984. Disponível em: https://doi.org/10.2307/2112465. Acesso em: 27 set. 2021.

[20] WINTERS, Katherine E.; MATUSOVICH, Holly M.; BRUNHAVER, Samantha R. **RECENT ENGINEERING GRADUATES MAKING CAREER CHOICES: FAMILY MATTERS**. Journal of Women and Minorities in Science and Engineering, v. 20, n. 4, p. 293-316, 2014. Disponível em: https://doi.org/10.1615/jwomenminorscieneng.2014008273. Acesso em: 27 set. 2021.

[21] SHUMBA, Almon; NAONG, Matsidiso. **Factors Influencing Students' Career Choice and Aspirations in South Africa**. Journal of Social Sciences, v. 33, n. 2, p. 169-178, nov. 2012. Disponível em: https://doi.org/10.1080/09718923.2012.11893096. Acesso em: 27 set. 2021.

[22] SIMMONS, Andrew N. **A Reliable Sounding Board: Parent Involvement in Students' Academic and Career Decision Making**. NACADA Journal, v. 28, n. 2, p. 33-43, 1 set. 2008. Disponível em: https://doi.org/10.12930/0271-9517-28.2.33. Acesso em: 27 set. 2021.

[23] JACOBSEN, Mary H. **Hand-Me-Down Dreams: How Families Influence Our Career Paths and How We Can Reclaim Them**. [S. l.]: Three Rivers Press, 2000. 240 p. ISBN 9780609802649.

[24] CHUANG, Ning-Kuang *et al*. **Hospitality Undergraduate Students' Career Choices and Factors Influencing Commitment to the Profession**. Journal of Hospitality &

Tourism Education, v. 19, n. 4, p. 28-37, out. 2007. Disponível em: https://doi.org/10.1080/10963758.2007.10696902. Acesso em: 27 set. 2021.

[25] OGHENETEGA, Timothy Oghenefega. **An exploratory study of attractors and detractors in Black graduates' choice of an academic career in a South African higher education institution**. 2017. Master Thesis — University of Cape Town, [s. l.], 2017. Disponível em: http://hdl.handle.net/11427/25414. Acesso em: 27 set. 2021.

[26] BRIGHT, Jim E. H.; PRYOR, Robert G. L.; HARPHAM, Lucy. **The role of chance events in career decision making**. Journal of Vocational Behavior, v. 66, n. 3, p. 561-576, jun. 2005. Disponível em: https://doi.org/10.1016/j.jvb.2004.05.001. Acesso em: 27 set. 2021.

[27] MCWHIRTER, Ellen Hawley; CROTHERS, Marciana; RASHEED, Saba. **The effects of high school career education on social-cognitive variables**. Journal of Counseling Psychology, v. 47, n. 3, p. 330-341, 2000. Disponível em: https://doi.org/10.1037/0022-0167.47.3.330. Acesso em: 27 set. 2021.

[28] **PSYCHOLOGY of mood and motivation**. In: INDIVIDUAL Differences and Personality. [S. l.]: Routledge, 2015. p. 199-216. ISBN 9780203785218. Disponível em: https://doi.org/10.4324/9780203785218-15. Acesso em: 27 set. 2021.

[29] PORTER, Stephen R.; UMBACH, Paul D. **COLLEGE MAJOR CHOICE: An Analysis of Person–Environment Fit. Research in Higher Education**, v. 47, n. 4, p. 429-449, 10 fev. 2006. Disponível em: https://doi.org/10.1007/s11162-005-9002-3. Acesso em: 27 set. 2021.

[30] HOLLAND, John L. **Making vocational choices: A theory of vocational personalities and work environments**. 2. ed. Odessa, Fla: Psychological Assessment Resources, 1992. 211 p. ISBN 091190705X.

[31] WAKAMATSU, Yosuke. **Type of career decision making process of undergraduates. The Proceedings of the Annual Convention of the Japanese Psychological Association**, v. 74, p. 2PM019, 20 set. 2010. Disponível em: https://doi.org/10.4992/pacjpa.74.0_2pm019. Acesso em: 27 set. 2021.

[32] BYKER SHANKS, Carmen *et al*. **Factors Influencing Food Choices Among Older Adults in the Rural Western USA**. Journal of Community Health, v. 42, n. 3, p. 511-521,

21 out. 2016. Disponível em: https://doi.org/10.1007/s10900-016-0283-6. Acesso em: 27 set. 2021.

[33] NYARKO – SAMPSON, E. **Teacher trainees' appraisal of guidance and counselling programmes in colleges of education in Ghana: A study of selected colleges in the Eastern and greater Accra zones**. Nigerian Journal of Guidance and Counselling, v. 15, n. 1, 22 mar. 2011. Disponível em: https://doi.org/10.4314/njgc.v15i1.64656. Acesso em: 27 set. 2021.

[34] BRIGHT, Jim E. H.; PRYOR, Robert G. L.; HARPHAM, Lucy. **The role of chance events in career decision making**. Journal of Vocational Behavior, v. 66, n. 3, p. 561-576, jun. 2005. Disponível em: https://doi.org/10.1016/j.jvb.2004.05.001. Acesso em: 27 set. 2021.

[35] BRIGHT, Jim E. H.; PRYOR, Robert G. L. **The Chaos Theory of Careers: A User's Guide. The Career Development Quarterly**, v. 53, n. 4, p. 291-305, jun. 2005. Disponível em: https://doi.org/10.1002/j.2161-0045.2005.tb00660.x. Acesso em: 27 set. 2021.

[36] SUUTARI, Vesa. **Global managers: career orientation, career tracks, life-style implications and career commitment**. Journal of Managerial Psychology, v. 18, n. 3, p. 185-207, maio 2003. Disponível em: https://doi.org/10.1108/02683940310465225. Acesso em: 27 set. 2021.

[37] **CAREER development: Issues of gender, race, and class**. Columbus, Ohio: ERIC Clearinghouse on Adult, Career, and Vocational Education, Center on Education and Training for Employment, College of Education, the Ohio State University, 1997.

[38] ISSA, AO; NWALO, KIN. **Influence Of Age, Gender, Subject Background And Predisposing Factors On The Admission Choice Of Undergraduates In Nigerian Library Schools**. Information Technologist (The), v. 5, n. 2, 17 fev. 2009. Disponível em: https://doi.org/10.4314/ict.v5i2.32030. Acesso em: 27 set. 2021.

[39] FREITAS, H. JANISSEK, R., MOSCAROLA, J. e BAULACc, Y. **Pesquisa interativa e novas tecnologias para coleta e análise de dados usando o Sphinx®**. Porto Alegre: Sphinx, 2002, 381p.

[40] GALAN, J.P.e VERNETTE, E. **Vers une 4èmè génération: les études de marché On-line**. França: Revue Décisions Marketing, n. 19, Jan-Abril 2000, pp.39-52.

[41] SCHONLAU, Matthias. **Conducting Research Surveys Via E-Mail and The Web**. [S. l.]: RAND Corporation, 2002. 112 p. ISBN 9780833031105.

[42] PITKOW, James E.; RECKER, Margaret M. **Using the Web as a survey tool: results from the second WWW user survey**. Computer Networks and ISDN Systems, v. 27, n. 6, p. 809-822, abr. 1995. Disponível em: https://doi.org/10.1016/0169-7552(95)00018-3. Acesso em: 27 set. 2021.

[43] JANISSEK, R. **A influência da Internet em negócios empresariais: identificação e caracterização de elementos para análise de sites**. Dissertação de Mestrado em Administração - Sistemas de Informação, PPGA/EA/UFRGS, Porto Alegre. Maio, 2000.

[44] BAULAC, Y., BOLDEN, R., MOSCAROLA, J. **Interactive Research: How Internet technology could revolutionise the survey and analysis process**. Londres: Association for Survey Computing Conference on Survey Research On The Internet, Imperial College, 28 Set. 2000.

[45] ISOTANI, S., & BITTENCOURT, I. I. **Dados Abertos Conectados: Em busca da Web do Conhecimento**. 2015. Novatec Editora.

[46] BEAVERS, Anthony F. Luciano Floridi: **Information: A Very Short Introduction. Minds and Machines**, v. 21, n. 1, p. 97-101, 28 jan. 2011. Disponível em: https://doi.org/10.1007/s11023-011-9224-4. Acesso em: 27 set. 2021.

[47] BITTENCOURT, I. I., ISOTANI, S., COSTA, E. & MIZOGUCHI, R. Research **Directions on Semantic Web and Education**. Journal Scientia, 19(1), 59-66. 2008.

[48] BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The Semantic Web**. Scientific American, v. 284, n. 5, p. 34-43, maio 2001. Disponível em: https://doi.org/10.1038/scientificamerican0501-34. Acesso em: 27 set. 2021.

[49] PÉRISSÉ, Marcelo Claudio. **SEMANTIC WEB IN HIGHER EDUCATION**. JISTEM Journal of Information Systems and Technology Management, v. 5, n. 2, p. 223-234, 1 ago. 2008. Disponível em: https://doi.org/10.4301/s1807-17752008000200002. Acesso em: 27 set. 2021.

[50] ISOTANI, Seiji *et al*. **A Semantic Web-based authoring tool to facilitate the planning of collaborative learning scenarios compliant with learning theories**. Computers & Education, v. 63, p. 267-284, abr. 2013. Disponível em: https://doi.org/10.1016/j.compedu.2012.12.009. Acesso em: 27 set. 2021.

[51] ISOTANI, Seiji; MIZOGUCHI, Riichiro. **Theory-Driven Group Formation through Ontologies**. In: ISOTANI, Seiji; MIZOGUCHI, Riichiro. Intelligent Tutoring Systems. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 646-655. ISBN 9783540691303. Disponível em: https://doi.org/10.1007/978-3-540-69132-7_67. Acesso em: 27 set. 2021.

[52] BITTENCOURT, I. I., ISOTANI, S., COSTA, E. & MIZOGUCHI, R. **Web 3.0-Os Rumos da Web Semântica e da Web 2.0 nos Ambientes Educacionais**. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBI*E).* (Vol. 1, No. 1, pp. 785-795). Nov. 2008.

[53] JANEV, Valentina; VRANEŠ, Sanja. **Applicability assessment of Semantic Web technologies**. Information Processing & Management, v. 47, n. 4, p. 507-517, jul. 2011. Disponível em: https://doi.org/10.1016/j.ipm.2010.11.002. Acesso em: 27 set. 2021.

[54] MIZOGUCHI, R.; BOURDEAU, J. **Theory-aware authoring environment-ontological engineering approach**. In: INTERNATIONAL CONFERENCE ON COMPUTERS IN EDUCATION, Auckland, New Zealand. International Conference on Computers in Education. [S. l.]: IEEE Comput. Soc. Disponível em: https://doi.org/10.1109/cie.2002.1186342. Acesso em: 27 set. 2021.

[55] JOVANOVIĆ, Jelena *et al*. **Leveraging the Social Semantic Web in Intelligent Tutoring Systems**. In: JOVANOVIĆ, Jelena *et al*. Intelligent Tutoring Systems. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 563-572. ISBN 9783540691303. Disponível em: https://doi.org/10.1007/978-3-540-69132-7_59. Acesso em: 27 set. 2021.

[56] DEVEDŽIĆ, Vladan. **Semantic Web and Education**. Boston, MA: Springer US, 2006. E-book. ISBN 9780387354163. Disponível em: https://doi.org/10.1007/978-0-387-35417-0. Acesso em: 27 set. 2021.

[57] BERNERS-LEE, Tim; HENDLER, James. **Scientific publishing on the 'semantic web'**. Nature, 12 abr. 2001. Disponível em: https://doi.org/10.1038/nature28055. Acesso em: 27 set. 2021.

[58] CHEATHAM, Michelle; PESQUITA, Catia. **Semantic Data Integration**. In: CHEATHAM, Michelle; PESQUITA, Catia. Handbook of Big Data Technologies. Cham: Springer International Publishing, 2017. p. 263-305. ISBN 9783319493398. Disponível em: https://doi.org/10.1007/978-3-319-49340-4_8. Acesso em: 27 set. 2021.

[59] HEATH, Tom; BIZER, Christian. **Linked Data: Evolving the Web into a Global Data Space**. Synthesis Lectures on the Semantic Web: Theory and Technology, v. 1, n. 1, p. 1-136, 9 fev. 2011. Disponível em: https://doi.org/10.2200/s00334ed1v01y201102wbe001. Acesso em: 27 set. 2021.

[60] HARMELEN, Frank Van *et al*. **Semantic Web Primer**. [S. l.]: MIT Press, 2012. 288 p. ISBN 9780262305617.

[61] PEREIRA, D. L. N. C. **Integração semântica das bases de dados do Sistema Único de Saúde: um estudo de caso com o Município de São Paulo**. Doctoral dissertation. Universidade de São Paulo. 2019.

[62] CRUZ, Isabel F. XIAO, Huiyong. **The role of ontologies in data integration**. International journal of engineering intelligent systems for electrical engineering and communications, 13(4):245–252, 2005.

[63] DOAN, AnHai; HALEVY, Alon; IVES, Zachary. Introduction. In: DOAN, AnHai; HALEVY, Alon; IVES, Zachary. **Principles of Data Integration**. [S. l.]: Elsevier, 2012. p. 1-18. ISBN 9780124160446. Disponível em: https://doi.org/10.1016/b978-0-12-416044-6.00001-6. Acesso em: 27 set. 2021.

[64] GROSSMANN, Wilfried; RINDERLE-MA, Stefanie. Introduction. In: GROSSMANN, Wilfried; RINDERLE-MA, Stefanie. **Data-Centric Systems and Applications**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015. p. 1-33. ISBN 9783662465301. Disponível em: https://doi.org/10.1007/978-3-662-46531-8_1. Acesso em: 27 set. 2021.

[65] CONSORTIUM, World Wide Web. **Document Object Model Level 1 Specification (Open Documents Standards Library)**. [S. l.]: Iuniverse Inc, 2000. 260 p. ISBN 9781583482544.

[66] BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data. In: BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. **Semantic Services, Interoperability and Web Applications**. [S. l.]: IGI Global, 2011. p. 205-227. ISBN 9781609605933. Disponível em: https://doi.org/10.4018/978-1-60960-593-3.ch008. Acesso em: 27 set. 2021.

[67] GRUBER, Thomas R. **A translation approach to portable ontology specifications**. Knowledge Acquisition, v. 5, n. 2, p. 199-220, jun. 1993. Disponível em: https://doi.org/10.1006/knac.1993.1008. Acesso em: 27 set. 2021.

[68] GARDNER, Stephen P. **Ontologies and semantic data integration**. Drug Discovery Today, v. 10, n. 14, p. 1001-1007, jul. 2005. Disponível em: https://doi.org/10.1016/s1359-6446(05)03504-x. Acesso em: 27 set. 2021.

[69] BUCCELLA, Agustina; CECHICH, Alejandra; FILLOTTRANI, Pablo. **Ontology-driven geographic information integration: A survey of current approaches**. Computers & Geosciences, v. 35, n. 4, p. 710-723, abr. 2009. Disponível em: https://doi.org/10.1016/j.cageo.2008.02.033. Acesso em: 27 set. 2021.

[70] KIRYAKOV, Atanas. Ontologies for Knowledge Management. In: KIRYAKOV, Atanas. **Semantic Web Technologies**. Chichester, UK: John Wiley & Sons, Ltd, 2006. p. 115-138. ISBN 9780470030332. Disponível em: https://doi.org/10.1002/047003033x.ch7. Acesso em: 27 set. 2021.

[71] CALVANESE, Diego *et al*. **Ontology-Based Data Access and Integration**. In: CALVANESE, Diego *et al*. Encyclopedia of Database Systems. New York, NY: Springer New York, 2018. p. 2590-2596. ISBN 9781461482666. Disponível em: https://doi.org/10.1007/978-1-4614-8265-9_80667. Acesso em: 27 set. 2021.

[72] KONTCHAKOV, Roman; RODRÍGUEZ-MURO, Mariano; ZAKHARYASCHEV, Michael. **Ontology-Based Data Access with Databases: A Short Course**. In: KONTCHAKOV, Roman; RODRÍGUEZ-MURO, Mariano; ZAKHARYASCHEV, Michael. Reasoning Web. Semantic Technologies for Intelligent Data Access. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 194-229. ISBN 9783642397837. Disponível em: https://doi.org/10.1007/978-3-642-39784-4_5. Acesso em: 27 set. 2021.

[73] XIAO, Guohui *et al*. Efficient Ontology-Based Data Integration with Canonical IRIs. In: XIAO, Guohui *et al*. **The Semantic Web**. Cham: Springer International Publishing, 2018. p. 697-713. ISBN 9783319934167. Disponível em: https://doi.org/10.1007/978-3-319-93417-4_45. Acesso em: 27 set. 2021.

[74] SABOU, Marta; EKAPUTRA, Fajar J.; BIFFL, Stefan. Semantic Web Technologies for Data Integration in Multi-Disciplinary Engineering. In: SABOU, Marta; EKAPUTRA, Fajar J.; BIFFL, Stefan. **Multi-Disciplinary Engineering for Cyber-Physical Production Systems**. Cham: Springer International Publishing, 2017. p. 301-329. ISBN 9783319563442. Disponível em: https://doi.org/10.1007/978-3-319-56345-9_12. Acesso em: 27 set. 2021.

[75] ALBERTS, Charl; MBALO, Ndileka F.; ACKERMANN, Christiaan J. **Adolescents' Perceptions of the Relevance of Domains of Identity Formation: A South African Cross-Cultural Study**. Journal of Youth and Adolescence, v. 32, n. 3, p. 169-184, jun. 2003. Disponível em: https://doi.org/10.1023/a:1022591302909. Acesso em: 27 set. 2021.

[76] HEATH, Tom; BIZER, Christian. **Linked Data: Evolving the Web into a Global Data Space**. Synthesis Lectures on the Semantic Web: Theory and Technology, v. 1, n. 1, p. 1-136, 9 fev. 2011. Disponível em: https://doi.org/10.2200/s00334ed1v01y201102wbe001. Acesso em: 27 set. 2021.

[77] ZAIDAN, Fernando Hadad; BAX, Marcello Peixoto. **Linked Open Data como forma de agregar valor às informações clínicas**. AtoZ: novas práticas em informação e conhecimento, v. 2, n. 1, p. 44, 18 ago. 2013. Disponível em: https://doi.org/10.5380/atoz.v2i1.41319. Acesso em: 27 set. 2021.

[78] ISOTANI, Seiji *et al*. **Estado da Arte em Web Semântica e Web 2.0: Potencialidades e Tendências da Nova Geração de Ambientes de Ensino na Internet**. Revista Brasileira de Informática na Educação, v. 17, n. 1, p. 30-42, jan. 2009. Disponível em: https://doi.org/10.5753/rbie.2009.17.01.30. Acesso em: 27 set. 2021.

[79] LAMMEL, Iuri; MIELNICZUK, Luciana. **Aplicação da Web Semântica no jornalismo**. Estudos em Jornalismo e Mídia, v. 9, n. 1, 5 jul. 2012. Disponível em: https://doi.org/10.5007/1984-6924.2012v9n1p180. Acesso em: 27 set. 2021.

[80] ROLIM, T. V., VIDAL, V. M. P., AVILA, C. V. S., CRUZ, M. M. L. D., BARRIO, M., & Queiroz, D. (2019, October). **Semanticsefaz: an ontology-based semantic portal for the government spending**. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 493-496).

[81] IBRAHIM, M. E. **An Ontology-based Hybrid Approach to Course Recommendation**. In Higher Education (Doctoral dissertation, University of Portsmouth). 2019.

[82] BERARDI, Rita; VIDAL, Vania; CASANOVA, Marco A. **R2BA - Rationalizing R2RML Mapping by Assertion**. In: 17TH INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 2015, Barcelona, Spain. 17th International Conference on Enterprise Information Systems. [S. l.]: SCITEPRESS - Science and and Technology Publications, 2015. Disponível em: https://doi.org/10.5220/0005337700050014. Acesso em: 27 set. 2021.

[83] HAZBER, Mohamed A. G. *et al*. **An Approach for Mapping Relational Database into Ontology**. In: 2015 12TH WEB INFORMATION SYSTEM AND APPLICATION CONFERENCE (WISA), 2015, Jinan, China. 2015 12th Web Information System and Application Conference (WISA). [S. l.]: IEEE, 2015. Disponível em: https://doi.org/10.1109/wisa.2015.25. Acesso em: 27 set. 2021.

[84] HAZBER, Mohamed A. G. *et al*. **An Approach for Automatically Generating R2RML-Based Direct Mapping from Relational Databases**. In: HAZBER, Mohamed A. G. *et al*. Communications in Computer and Information Science. Singapore: Springer Singapore, 2016. p. 151-169. ISBN 9789811020520. Disponível em: https://doi.org/10.1007/978-981-10-2053-7_15. Acesso em: 27 set. 2021.

Anexo 01 – Código fonte das TR's

```
CREATE TABLE public.provenance_statement (
id int4 NOT NULL DEFAULT nextval('source_data_id_seq'::regclass),
title varchar(150) NULL,
url varchar(255) NULL,
last_extraction timestamptz(0) NULL,
"source" varchar(20) NULL,
codigo varchar(255) NULL,
populated bool NULL,
modified timestamptz(0) NULL,
creator varchar(255) NULL,
created timestamptz(0) NULL,
CONSTRAINT source_data_pk PRIMARY KEY (id),
CONSTRAINT source_data_un UNIQUE (codigo)
);

CREATE TABLE public.trabalho (
id serial NOT NULL,
titulo varchar(255) NULL,
descricao varchar(2048) NULL,
requisitos varchar(255) NULL,
remuneracao varchar(20) NULL,
localizacao varchar(255) NULL,
modo varchar(255) NULL,
area_curso varchar(255) NULL,
provenance_statement_id int2 NULL,
CONSTRAINT perfil_trabalho_pk PRIMARY KEY (id),
CONSTRAINT trabalho_fk FOREIGN KEY (provenance_statement_id)
REFERENCES public.provenance_statement(id)
);

CREATE TABLE public.college_or_university (
```

```
	id int4 NOT NULL DEFAULT
nextval('instituicoes_do_ensino_superior_id_seq'::regclass),
	"RowKey" varchar(255) NULL,
	"Timestamp" timestamptz NULL,
	entityid varchar(255) NULL,
	codigodoestabelecimento varchar(255) NULL,
	nomedoestabelecimento varchar(255) NULL,
	morada varchar(255) NULL,
	codigopostal varchar(255) NULL,
	distrito varchar(255) NULL,
	concelho varchar(255) NULL,
	tipodeensino varchar(255) NULL,
	paginaweb varchar(255) NULL,
	email varchar(255) NULL,
	telefone varchar(255) NULL,
	fax varchar(255) NULL,
	"PartitionKey" varchar(255) NULL,
	codigodependede varchar(255) NULL,
	dependede varchar(255) NULL,
	provenance_statement_id int2 NULL,
	link_open_data varchar NULL,
	CONSTRAINT instituicoes_do_ensino_superior_estabelecimento_un UNIQUE
(codigodoestabelecimento),
	CONSTRAINT instituicoes_do_ensino_superior_pkey PRIMARY KEY (id),
	CONSTRAINT instituicoes_do_ensino_superior_un UNIQUE ("RowKey",
"PartitionKey", provenance_statement_id),
	CONSTRAINT instituicoes_do_ensino_superior_fk FOREIGN KEY
(provenance_statement_id) REFERENCES public.provenance_statement(id)
	);

	CREATE TABLE public.curso_cnaef (
	id int4 NOT NULL DEFAULT
nextval('class_nacional_de_areas_de_educacao_e_formacao_id_seq'::regclass),
```

```
	"RowKey" varchar(255) NULL,
	"Timestamp" timestamptz NULL,
	entityid varchar(255) NULL,
	subsistema varchar(255) NULL,
	tipodeensino varchar(255) NULL,
	distrito varchar(255) NULL,
	estabelecimento varchar(255) NULL,
	nome varchar(255) NULL,
	niveldeformacao varchar(255) NULL,
	areacnaef varchar(255) NULL,
	"PartitionKey" varchar(255) NULL,
	provenance_statement_id int2 NULL,
	CONSTRAINT class_nacional_de_areas_de_educacao_e_formacao_pkey PRIMARY
KEY (id),
	CONSTRAINT class_nacional_de_areas_de_educacao_e_formacao_un UNIQUE
("RowKey", "PartitionKey", provenance_statement_id),
	CONSTRAINT class_nacional_de_areas_de_educacao_e_foptobvrmacao_fk
FOREIGN KEY (provenance_statement_id) REFERENCES public.provenance_statement(id),
	CONSTRAINT class_nacional_de_areas_de_educacao_e_formacao_fk FOREIGN
KEY (estabelecimento) REFERENCES
public.college_or_university(codigodoestabelecimento)
	);

	CREATE TABLE public.curso (
	id serial NOT NULL,
	url varchar(255) NULL,
	descricao varchar(4096) NULL,
	valor_propina_nacional int4 NULL,
	valor_propina_internacional int4 NULL,
	duracao varchar(50) NULL,
	modo varchar(255) NULL,
	curso_cnaef_id int4 NULL,
	provenance_statement_id int4 NULL,
```

```
        CONSTRAINT perfil_curso_pk PRIMARY KEY (id),
        CONSTRAINT curso_fk FOREIGN KEY (curso_cnaef_id) REFERENCES
public.curso_cnaef(id),
        CONSTRAINT curso_provenance_fk FOREIGN KEY (provenance_statement_id)
REFERENCES public.provenance_statement(id)
        );
```

Anexo 02 – Código fonte do extrator da UTAD

```
class UTADSpider(scrapy.Spider):
    name = "UTAD_spider"
    start_urls = ['https://www.utad.pt/estudar/inicio/licenciaturas-mestrados-integrados/', 'https://www.utad.pt/estudar/inicio/licenciaturas-mestrados-integrados/page/2/']
    instituicao_id = 1  # important
    valor_anual_nacional = 697
    valor_anual_internacional = 1500
    data_raspagem = datetime.datetime.now()
    fields = [
        'instituicao_id',
        'nome',
        'qualificacao',
        'url',
        'descricao',
        'campo_estudo',
        'area',
        'valor_anual_nacional',
        'valor_anual_internacional',
        'duracao',
        'modo',
        'data_raspagem'
    ]

    def parse(self, response):
        links_cursos = response.css('.linklist ::attr(href)')
        yield from response.follow_all(links_cursos, self.parse_curso)

    def parse_curso(self, response):

        def extract_with_css(query):
```

```
                return response.css(query).get(default='').strip()

            def extract_list_with_css(query):
                return response.css(query)

            nome = extract_with_css('h1.entry-title ::text')
            dados_curso = extract_list_with_css('.pe-tabs .row')
            url = response.url
            qualificacao = dados_curso[0].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '')
            descricao = dados_curso[1].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '')
            area = dados_curso[2].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '').split(' <a ')[0]
            campo_estudo = dados_curso[2].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '').split(' <a ')[0]
            duracao = dados_curso[6].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '')
            modo = dados_curso[5].css('div .col-xs-
12').extract_first(default='').replace('<br>', '').replace('</div>', '').replace('<div class="col-
xs-12 col-md-8">', '')

            values = [
                self.instituicao_id,
                nome,
                qualificacao,
                url,
                descricao,
```

```
                campo_estudo,
                area,
                self.valor_anual_nacional,
                self.valor_anual_internacional,
                duracao,
                modo,
                self.data_raspagem
            ]

            insertDB(connection=connectDB(), tabela='perfil_curso', fields=self.fields,
values=values)
```

Anexo 03 – Código fonte do extrator do *website* "net-empregos.com"

```
import datetime
import scrapy
from db_tools import *
from urllib.parse import unquote
import json
from io import StringIO
from html.parser import HTMLParser

class MLStripper(HTMLParser):
    def __init__(self):
        super().__init__()
        self.reset()
        self.strict = False
        self.convert_charrefs= True
        self.text = StringIO()
    def handle_data(self, d):
        self.text.write(d)
    def get_data(self):
        return self.text.getvalue()

def strip_tags(html):
    s = MLStripper()
    s.feed(html)
    return s.get_data()

def cleanup(url):
    try:
        return unquote(url, errors='strict')
    except UnicodeDecodeError:
        return unquote(url, encoding='latin-1')
```

```
class NETEMPREGOSpider(scrapy.Spider):
    name = "NETEMPREGO_spider"

    option_dict = {'29': 'Administração / Secretariado', '39': 'Agricultura /
Florestas / Pescas', '22': 'Arquitectura / Design', '40': 'Artes / Entretenimento / Media',
'16': 'Banca / Seguros / Serviços "Financeiros',
                   '47': 'Beleza / Moda / Bem Estar', '57': 'Call Center / Help Desk', '53':
'Comercial / Vendas', '8': 'Comunicação Social / Media', '51': 'Conservação /
Manutenção / Técnica', '23': 'Construção Civil',
                   '15': 'Contabilidade / Finanças',
                   '28': 'Desporto / Ginásios', '44': 'Direito / Justiça', '11': 'Educação /
Formação', '54': 'Engenharia ( Ambiente )', '45': 'Engenharia ( Civil )', '46': 'Engenharia (
Eletrotecnica )', '24': 'Engenharia ( Mecanica )',
                   '50': 'Engenharia ( Química / Biologia )', '41': 'Farmácia /
Biotecnologia', '26': 'Gestão de Empresas / Economia', '32': 'Gestão RH', '9': 'Hotelaria /
Turismo', '12': 'Imobiliário', '6': 'Indústria / Produção',
                   '38': 'Informática ( Analise de Sistemas )', '34': 'Informática (
Formação )', '37': 'Informática ( Gestão de Redes )', '35': 'Informática ( Internet )', '36':
'Informática ( Multimedia )', '5': 'Informática ( Programação )',
                   '49': 'Informática ( Técnico de Hardware )', '56': 'Informática
(Comercial/Gestor de Conta)', '58': 'Limpezas / Domésticas', '30': 'Lojas / Comércio /
Balcão', '19': 'Publicidade / Marketing', '18': 'Relações Públicas',
                   '42': 'Restauração / Bares / Pastelarias', '14': 'Saúde / Medicina /
Enfermagem', '55': 'Serviços Sociais', '52': 'Serviços Técnicos', '1': 'Telecomunicações',
'43': 'Transportes / Logística'}

    area_dict = {}

    data_raspagem = datetime.datetime.now()

    fields = [
        'titulo',
```

```
            'descricao',
            'requisitos',
            'remuneracao',
            'localizacao',
            'modo',
            'data_raspagem'
        ]

        # link base: "Formação Superior", "Tempo Integral"
        link_base = 'https://www.net-empregos.com/pesquisa-
empregos.asp?chaves=Forma%E7%E3o+Superior&cidade=&categoria={}&zona=0&ti
po=1'

        areas = selectDB(connectDB(), 'select area from perfil_curso')

        start_urls = []
        all_areas_list = []
        for x in areas:
            y = x[0].split(' ')
            base_world = []
            for i in y:
                if len(i) > 3:
                    base_world.append(i)
            print('base_world', base_world)
            option_list = []
            for i in base_world:
                for key, value in option_dict.items():
                    if i in value:
                        option_list.append(key)
            print('option_list', option_list)
            option_list = list(set(option_list)) # retira duplicados
            area_dict[x[0]] = option_list
            all_areas_list = all_areas_list + option_list
```

```
            all_areas_list = list(set(all_areas_list))
            for i in all_areas_list:
                start_urls.append(link_base.format(i))

            def parse(self, response):

                links_jobs = response.css('.div-right a ::attr(href)')

                yield from response.follow_all(links_jobs, self.parse_job)

                next_page = response.css('a .page-link .d-none .d-lg-block
::attr(href)').extract_first()
                print('next_page', next_page)
                if next_page and next_page not in self.urls:
                    yield scrapy.Request(
                        response.urljoin(next_page),
                        callback = self.parse
                    )

            def parse_job(self, response):

                def extract_with_css(response, query):
                    return response.css(query).extract_first(default='')

                def extract_list_with_css(response, query):
                    return response.css(query)

                link = extract_list_with_css(response, '.job-details-page')
                for x in link:
                    titulo = extract_with_css(x, 'h1 ::text')
                    print('TITLE', titulo)
                    description = extract_list_with_css(x, 'p')
```

```
                for y in description:
                    # print(strip_tags(y))
                    _y = extract_list_with_css(y, '::text')
                    descricao = ''
                    for k in _y:
                        descricao = descricao + k.get().lstrip().rstrip() + ' '

                localizacao = 'Veja em Descrição'
                modo = 'Veja em Descrição'

                categoria = cleanup(str(response.request.headers['Referer'].decode("utf-
8")).split('categoria=')[1].split('&zona')[0].replace('+', ' '))

                remuneracao = 'Não informada'

                if titulo:
                    for key, value in self.area_dict.items():
                        if categoria in value:
                            requisitos = key
                            insertDB(connection=connectDB(), tabela='perfil_trabalho',
fields=self.fields, values=[titulo, descricao, requisitos, remuneracao, localizacao, modo,
self.data_raspagem])
```

Anexo 04 – Código fonte da OP

```
@prefix : <https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp#> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix psesp: <https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#> .
@prefix schema: <http://schema.org/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ontologia_psesp: <https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp#> .
@base <https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp> .
<https://github.com/LuSoMaBra/semantic-esp/tree/master/semanticesp/ontologia_psesp> rdf:type owl:Ontology .
#################################################################
#    Datatypes
#################################################################
###  http://purl.org/dc/terms/creator
dcterms:creator rdf:type rdfs:Datatype .
###  http://schema.org/addressLocality
schema:addressLocality rdf:type rdfs:Datatype .
###  http://schema.org/branchCode
schema:branchCode rdf:type rdfs:Datatype .
###  http://schema.org/description
schema:description rdf:type rdfs:Datatype .
###  http://schema.org/jobTitle
schema:jobTitle rdf:type rdfs:Datatype .
```

```
        ###  http://schema.org/name
        schema:name rdf:type rdfs:Datatype .
        ###  http://schema.org/title
        schema:title rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#AreaCurso
        psesp:AreaCurso rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#DuracaoCurso
        psesp:DuracaoCurso rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ModoTrabalho
        psesp:ModoTrabalho rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#RemuneracaoTrabalho
        psesp:RemuneracaoTrabalho rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ValorPropinaInternacional
        psesp:ValorPropinaInternacional rdf:type rdfs:Datatype .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#source
        psesp:source rdf:type rdfs:Datatype .
        #################################################################
        #    Object Properties
        #################################################################
        ###  http://purl.org/dc/terms/created
        dcterms:created rdf:type owl:ObjectProperty ;
                        rdfs:domain dcterms:ProvenanceStatement ;
                        rdfs:range xsd:datetime .
        ###  http://purl.org/dc/terms/hasPart
        dcterms:hasPart rdf:type owl:ObjectProperty ;
                        rdfs:domain dcterms:ProvenanceStatement ;
                        rdfs:range schema:CollegeOrUniversity ,
```

```
                        schema:Course ,
                        psesp:Trabalho ;
                rdfs:label "tem parte de"@pt .
        ###  http://purl.org/dc/terms/isPartOf
        dcterms:isPartOf rdf:type owl:ObjectProperty ;
                rdfs:domain schema:CollegeOrUniversity ;
                rdfs:range schema:Course ;
                rdfs:label "é parte de"@pt .
        ###  http://purl.org/dc/terms/modified
        dcterms:modified rdf:type owl:ObjectProperty ;
                rdfs:domain dcterms:ProvenanceStatement ;
                rdfs:range xsd:datetime .
        ###  http://schema.org/sameAs
        schema:sameAs rdf:type owl:ObjectProperty ;
                rdfs:domain psesp:RequisitosTrabalho .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#isPropertyOff
        psesp:isPropertyOff rdf:type owl:ObjectProperty ;
                rdfs:domain dcterms:ProvenanceStatement ,
                        schema:CollegeOrUniversity ,
                        schema:Course ,
                        psesp:Trabalho ;
                rdfs:range dcterms:created ,
                        dcterms:modified ,
                        schema:addressRegion ,
                        schema:courseMode ,
                        schema:postalCode ,
                        schema:streetAddress ,
                        schema:url ,
                        psesp:AreaCNAEF ,
                        psesp:NivelDeFormacao ,
                        psesp:ValorPropinaNacional ,
                        psesp:lastExtraction .
```

```
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#lastExtraction
        psesp:lastExtraction rdf:type owl:ObjectProperty ;
                        rdfs:domain dcterms:ProvenanceStatement ;
                        rdfs:range xsd:datetime .
        ###############################################################################
        #    Data properties
        ###############################################################################
        ###  http://purl.org/dc/terms/created
        dcterms:created rdf:type owl:DatatypeProperty .
        ###  http://purl.org/dc/terms/creator
        dcterms:creator rdf:type owl:DatatypeProperty ;
                    rdfs:domain dcterms:ProvenanceStatement ;
                    rdfs:range xsd:string .
        ###  http://purl.org/dc/terms/modified
        dcterms:modified rdf:type owl:DatatypeProperty .
        ###  http://schema.org/addressLocality
        schema:addressLocality rdf:type owl:DatatypeProperty ;
                         rdfs:domain schema:PostalAddress ;
                         rdfs:range xsd:string .
        ###  http://schema.org/addressRegion
        schema:addressRegion rdf:type owl:DatatypeProperty ;
                        rdfs:domain schema:CollegeOrUniversity ;
                        rdfs:range xsd:string .
        ###  http://schema.org/branchCode
        schema:branchCode rdf:type owl:DatatypeProperty ;
                    rdfs:domain schema:CollegeOrUniversity ;
                    rdfs:range xsd:string .
        ###  http://schema.org/courseMode
        schema:courseMode rdf:type owl:DatatypeProperty ;
                    rdfs:domain schema:Course ;
                    rdfs:range xsd:string .
        ###  http://schema.org/description
```

```
schema:description rdf:type owl:DatatypeProperty ;
                   rdfs:domain schema:Course ,
                               psesp:Trabalho ;
                   rdfs:range xsd:string .
###  http://schema.org/jobTitle
schema:jobTitle rdf:type owl:DatatypeProperty ;
                rdfs:domain psesp:Trabalho ;
                rdfs:range xsd:string .
###  http://schema.org/name
schema:name rdf:type owl:DatatypeProperty ;
            rdfs:domain schema:CollegeOrUniversity ,
                        schema:Course ;
            rdfs:range xsd:string .
###  http://schema.org/postalCode
schema:postalCode rdf:type owl:DatatypeProperty ;
                  rdfs:domain schema:CollegeOrUniversity ;
                  rdfs:range xsd:string .
###  http://schema.org/sameAs
schema:sameAs rdf:type owl:DatatypeProperty ;
              rdfs:range psesp:AreaCurso .
###  http://schema.org/streetAddress
schema:streetAddress rdf:type owl:DatatypeProperty ;
                     rdfs:domain schema:CollegeOrUniversity ;
                     rdfs:range xsd:string .
###  http://schema.org/title
schema:title rdf:type owl:DatatypeProperty ;
             rdfs:domain dcterms:ProvenanceStatement ;
             rdfs:range xsd:string .
###  http://schema.org/url
schema:url rdf:type owl:DatatypeProperty ;
           rdfs:domain dcterms:ProvenanceStatement ,
                       schema:CollegeOrUniversity ,
                       schema:Course ;
```

```
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#AreaCNAEF
        psesp:AreaCNAEF rdf:type owl:DatatypeProperty ;
                rdfs:domain schema:Course ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#AreaCurso
        psesp:AreaCurso rdf:type owl:DatatypeProperty ;
                rdfs:domain schema:Course ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#DuracaoCurso
        psesp:DuracaoCurso rdf:type owl:DatatypeProperty ;
                rdfs:domain schema:Course ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ModoTrabalho
        psesp:ModoTrabalho rdf:type owl:DatatypeProperty ;
                rdfs:domain psesp:Trabalho ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#NivelDeFormacao
        psesp:NivelDeFormacao rdf:type owl:DatatypeProperty ;
                rdfs:domain schema:Course ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#RemuneracaoTrabalho
        psesp:RemuneracaoTrabalho rdf:type owl:DatatypeProperty ;
                rdfs:domain psesp:Trabalho ;
                rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#RequisitosTrabalho
```

```
        psesp:RequisitosTrabalho rdf:type owl:DatatypeProperty ;
                        rdfs:domain psesp:Trabalho ;
                        rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ValorPropinaInternacional
        psesp:ValorPropinaInternacional rdf:type owl:DatatypeProperty ;
                            rdfs:domain schema:CollegeOrUniversity ;
                            rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ValorPropinaNacional
        psesp:ValorPropinaNacional rdf:type owl:DatatypeProperty ;
                            rdfs:domain schema:CollegeOrUniversity ;
                            rdfs:range xsd:string .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#isPropertyOff
        psesp:isPropertyOff rdf:type owl:DatatypeProperty ;
                        rdfs:range dcterms:creator ,
                                schema:addressLocality ,
                                schema:branchCode ,
                                schema:description ,
                                schema:jobTitle ,
                                schema:name ,
                                schema:title ,
                                psesp:AreaCurso ,
                                psesp:DuracaoCurso ,
                                psesp:ModoTrabalho ,
                                psesp:RemuneracaoTrabalho ,
                                psesp:RequisitosTrabalho ,
                                psesp:ValorPropinaInternacional ,
                                psesp:source .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#lastExtraction
        psesp:lastExtraction rdf:type owl:DatatypeProperty .
```

```
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#source
        psesp:source rdf:type owl:DatatypeProperty ;
                rdfs:domain dcterms:ProvenanceStatement ;
                rdfs:range xsd:string .
        ###############################################################################
        #    Classes
        ###############################################################################
        ###  http://purl.org/dc/terms/ProvenanceStatement
        dcterms:ProvenanceStatement rdf:type owl:Class ;
                              rdfs:label "Provedor da Informação"@pt .
        ###  http://schema.org/CollegeOrUniversity
        schema:CollegeOrUniversity rdf:type owl:Class ;
                            rdfs:label "Instituição de Curso Superior"@pt .
        ###  http://schema.org/Course
        schema:Course rdf:type owl:Class ;
                 rdfs:label "Curso Superior"@pt .
        ###  http://schema.org/PostalAddress
        schema:PostalAddress rdf:type owl:Class .
        ###  http://schema.org/addressRegion
        schema:addressRegion rdf:type owl:Class .
        ###  http://schema.org/courseMode
        schema:courseMode rdf:type owl:Class .
        ###  http://schema.org/postalCode
        schema:postalCode rdf:type owl:Class .
        ###  http://schema.org/streetAddress
        schema:streetAddress rdf:type owl:Class .
        ###  http://schema.org/url
        schema:url rdf:type owl:Class .
        ###  http://www.w3.org/2001/XMLSchema#datetime
        xsd:datetime rdf:type owl:Class .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#AreaCNAEF
```

```
        psesp:AreaCNAEF rdf:type owl:Class .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#NivelDeFormacao
        psesp:NivelDeFormacao rdf:type owl:Class .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#Trabalho
        psesp:Trabalho rdf:type owl:Class ;
                rdfs:label "Oferta de Trabalho"@pt .
        ###  https://github.com/LuSoMaBra/semantic-
esp/tree/master/semanticesp/ontology/ontologia_psesp.owl#ValorPropinaNacional
        psesp:ValorPropinaNacional rdf:type owl:Class .
        #################################################################
        #    Annotations
        #################################################################
        dcterms:created rdfs:label "Data e hora da criação da informação"@pt .
        dcterms:creator rdfs:label "Nome do criador da informação"@pt .
        dcterms:modified rdfs:label "Data e hora da última modificação da
informação"@pt .
        schema:addressLocality rdfs:label "Concelho"@pt .
        schema:addressRegion rdfs:label "Distrito"@pt .
        schema:branchCode rdfs:label "Código da Universidade ou Instituto de ensino
superior"@pt .
        schema:courseMode rdfs:label "Modo curso (Presencial/Remoto)"@pt .
        schema:description rdfs:label "Descrição"@pt .
        schema:jobTitle rdfs:label "Título da oferta de emprego"@pt .
        schema:name rdfs:label "Área profissional do curso"@pt ,
                        "Nome da Universidade ou Instituto de ensino superior"@pt .
        schema:postalCode rdfs:label "Código postal"@pt .
        schema:sameAs rdfs:label "é igual a"@pt .
        schema:streetAddress rdfs:label "Morada"@pt .
        schema:title rdfs:label "Nome do provedor da informação"@pt .
        schema:url rdfs:label "URL de origem da informação"@pt .
        psesp:AreaCNAEF rdfs:label "Área profissional do curso"@pt .
```

```
        psesp:AreaCurso rdfs:label "Área de estudo do curso"@pt .
        psesp:DuracaoCurso rdfs:label "Duração do curso (anos)"@pt .
        psesp:ModoTrabalho rdfs:label "Modo de trabalho
(remoto/presencial/ambos)"@pt .
        psesp:NivelDeFormacao rdfs:label "Qualificação do curso"@pt .
        psesp:RemuneracaoTrabalho rdfs:label "Remuneração oferecidade pela oferta de
emprego"@pt .
        psesp:RequisitosTrabalho rdfs:label "Requisitos académicos para ocupar a vaga
de emprego (Área do Curso)"@pt .
        psesp:ValorPropinaInternacional rdfs:label "Valor da propina para alunos
internacionais"@pt .
        psesp:ValorPropinaNacional rdfs:label "Valor da propina para alunos
nacionais"@pt .
        psesp:isPropertyOff rdfs:label "é propriedade de"@pt .
        psesp:lastExtraction rdfs:label "Data e hora da extração da informação"@pt .
        psesp:source rdfs:label "Formato da informação"@pt .
        ###  Generated by the OWL API (version 4.5.9.2019-02-01T07:24:44Z)
https://github.com/owlcs/owlapi
```