Universidade do Minho
Escola de Engenharia
Departamento de Informática

José Pedro Saraiva de Carvalho

Uma Ontologia para o Livro das Propriedades da Mesa Arcebispal de Braga

December 2021

Universidade do Minho
Escola de Engenharia
Departamento de Informática

José Pedro Saraiva de Carvalho

Uma Ontologia para o Livro das Propriedades da Mesa Arcebispal de Braga

Master dissertation
Integrated Master's in Informatics Engineering

Dissertation supervised by
Professor Orlando Manuel de Oliveira Belo
Professora Anabela Leal de Barros

December 2021

AGRADECIMENTOS

Antes do começo desta dissertação é necessário mencionar algumas pessoas que tiveram papeis cruciais no seu desenvolvimento. Primeiramente, gostaria de agradecer ao meu orientador, o Professor Doutor Orlando Manuel Oliveira Belo, à minha coorientadora a Professora Doutora Anabela Leal Barros por toda a ajuda e pela grande quantidade de tempo que disponibilizaram para me auxiliar e aconselhar ao longo do percurso de desenvolvimento desta dissertação.

Gostaria também de agradecer aos meus colegas Tiago Fraga, João Gomes e Ricardo Martins pelo auxílio que prestaram e pelas explicações que deram acerca do projeto, que permitiram acelerar a sua realização.

De seguida gostaria de agradecer aos meus pais e ao meu irmão, por toda a ajuda e apoio que me prestaram, e por todo o amor que me deram ao longo da minha vida e em especial ao longo do desenvolvimento desta dissertação, sem o qual não teria sido possível eu ter tido a resiliência para a concluir.

A todas as outras pessoas, quer sejam outros familiares ou amigos, dentro e fora da universidade, obrigado por me terem ajudado a forjar o meu percurso académico e por me terem apoiado de forma a eu poder ter alcançado este patamar. Não teria sido possível sem todos eles.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

ABSTRACT

With the development of several areas of human knowledge comes an increase in the quantity of information that is required to be stored. It is therefore necessary to supply some structure to this information and to organize it in a way that it becomes more explicit and presented in a less dense format, making it easier to be consulted and analised. To find a solution to this problem it is normal to resort to the construction of ontologies. These can be described as specifications of domain conceptualizations. In other words they try to describe, primarily through the use of concepts, relationships between them and rules or axioms, the structure of knowledge domain. The level of organization that ontologies supply to any domain that is by them explained makes them usefull in several scenarios, which causes them to be applied in very different settings and areas of knowledge regularly. Every component of an ontology is an intrical part to the truthfull representation of the data it contains. Due to this fact, it is necessary that the construction of an ontology corresponds to a meticulous and well-organized process in order not to represent corrupt or incorrect data regarding a domain. The search for continually improving versions of these types of processes has resulted in several approaches to the building of an ontology, which has culminated in the development of systems with the purpose of aiding in the several steps of this process. Even though these systems exist, the vast majority of ontologies are still elaborated manually, which can affect the created ontology in a very negative way, that can vary from time it takes to complete the process of its creation to the degree of success in which it is completed. This dissertation will discuss the semiautomatic process of bulding an ontology representative of the Tombo da Mitra Arquiepiscopal de Braga, with the goal of organizing and facilitating the comprehension of the information kept within this codex of the early XVII century.

KEYWORDS Ontologies, Ontolog Learning, Semiautomatic Building of Ontologies, Natural Language Processing, Lexical-Syntactic Patterns, Tombo da Mitra.

RESUMO

Com o desenvolvimento de várias áreas do saber humano, a quantidade de informação que é necessário armazenar aumenta de forma considerável dia após dia. Este facto torna necessária a organização e estruturação desta informação, de forma que esta fique mais explícita e apresentada num formato menos denso, para que a sua consulta e análise sejam facilitadas. Para solucionar este problema, atualmente recorre-se à elaboração de ontologias. Estas podem ser descritas como especificações de conceptualizações relativas a domínios específicos, que descrevem, primariamente através de conceitos, relações entre estes e regras/axiomas, a estrutura de um domínio de conhecimento. O modelo de organização da informação que as ontologias disponibilizam faz com que sejam utilizadas nos mais variados domínios e cenários aplicacionais, promovendo um leque de aplicações bastante vasto e alargando a sua área de aplicação. Cada um dos componentes de uma ontologia é vital para a veracidade da informação representada. Devido a este facto, é necessário que a elaboração de uma ontologia seja suportada por um processo bem organizado e meticuloso, por forma a não representar informação errónea ou de forma incorreta. A procura por versões cada vez melhores deste tipo de processos resultou em várias abordagens à construção de uma ontologia, o que resultou no desenvolvimento de sistemas próprios cujo propósito é auxiliar nos diversos passos deste processo. Apesar da existência destes sistemas, a grande maioria das ontologias continua a ser elaborada de forma manual, o que provoca várias anomalias na construção da própria ontologia, quer no tempo que este processo pode levar, quer na substância do seu conteúdo, quando o processo se encontra concluído. Nesta dissertação será abordado o processo semiautomático para a elaboração de uma ontologia representativa do conteúdo do Tombo da Mitra Arquiepiscopal de Braga, com o intuito de estruturar e de facilitar a compreensão da informação contida neste códice de inícios do século XVII.

PALAVRAS-CHAVE Ontologias, Aprendizagem de Ontologias, Elaboração Semiautomática de Ontologias, Processamento de Linguagem Natural, Padrões léxico-sintáticos, Tombo da Mitra.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

Parte I

ESTUDO INTRODUTÓRIO

1

INTRODUÇÃO

1.1 ENQUADRAMENTO GERAL

A perceção da realidade que nos circunda é algo que o ser humano estuda e em que tem interesse, sendo esta acompanhada da tentativa da sua representação e classificação dos seus elementos o que culminou no desenvolvimento do termo *ontologia*.

Apesar disso é necessário referir que uma ontologia não é descrita como uma mera classificação de informação, uma vez que representa um dado domínio de conhecimento de forma muito mais rica, estabelecendo relações entre os seus elementos constituintes, categorizando-os, ao invés de uma classificação que, simplesmente, representa uma categorização de diferentes conceitos (Rees 2003).

A definição de *ontologia* vem normalmente acompanhada de duas perspetivas distintas: uma proveniente da Filosofia e outra proveniente da Informática. O estudo de ontologias nestas duas áreas varia em termos de propósito, obviamente. De entre estas duas, a definição de *ontologia* mais antiga é a da Filosofia, tendo esta mais de 2400 anos de idade. Esta perspetiva foca-se na análise categórica da realidade, ou seja, na descoberta das entidades existentes e das diferentes categorias de entidades a que estas pertencem, com um intuito de elaborar um inventário da realidade (Poli, Healy e Kameas 2010). Na perspetiva da Filosofia, uma ontologia é definida por Huang (2010) como a ciência do que é, dos tipos e estruturas de objetos, propriedades, eventos, processos e relações em cada área da realidade.

A perspetiva proveniente da Informática, por vezes referida como *Engenharia de Ontologias*, é bastante mais recente, tendo apenas algumas décadas de idade. Esta coloca as mesmas questões que a abordagem apresentada na Filosofia, tendo, no entanto, intenções bastante diferentes. As intenções são segundo Poli, Healy e Kameas (2010) *"to create engineering models of reality, artifacts which can be used by software, and perhaps directly interpreted and reasoned over by special software called inference engines, to imbue software with human level semantics"*, ou seja, representar a realidade de forma que esta possa ser interpretada por *software*.

O conhecimento passível de ser obtido a partir da realidade pode ser dividido em diversas áreas ou domínios do saber, podendo cada uma destas ser representada pela sua própria ontologia. Até recentemente, grande parte das ontologias era criada e organizada de forma manual, podendo ser utilizado nesse processo algum *software* de representação de conhecimento, como, por exemplo o Protégé (*Protégé*).

Porém a elaboração de ontologias de forma manual originava alguns erros, frequentemente provenientes da condição humana, e que iam assumindo proporções cada vez mais grotescas à medida que a quantidade de

informação do domínio de trabalho ia aumentando – uma situação que, com sabemos, é cada vez mais comum. Este é um problema de grande relevância, especialmente com o papel cada vez mais preponderante que as ontologias vão tendo em várias áreas de aplicação.

Para resolver este problema foram elaborados ao longo dos últimos anos sistemas cujo propósito era obter ontologias a partir de textos não estruturados, tentando minimizar a intervenção humana nos seus procedimentos. Estes sistemas não se limitaram a um tipo de abordagem ou modelo, utilizando, geralmente, uma combinação destes. Cada um dos componentes é somente utilizado na área na qual se consegue retornar melhores resultados.

Estes sistemas também foram sofrendo uma evolução em termos da tecnologia que utilizavam, efetuando na última década a transição de abordagens de *Shallow Learning* para abordagens mais centradas em modelos de *Deep Learning*.

1.2 ELABORAÇÃO (SEMI)AUTOMÁTICA DE ONTOLOGIAS

Atualmente já existem metodologias desenhadas com o intuito de otimizar e fazer com que a elaboração de ontologias de forma manual seja bem efetuada. Porém apesar da existência deste tipo de estratégias estar bem delineada, algumas áreas do saber possuem dimensões tão vastas que a otimização deste tipo de metodologias não torna viável a elaboração manual de ontologias. Numa tentativa de resolução deste tipo de problemas muitos engenheiros do conhecimento começaram a tentar automatizar algumas etapas deste processo. Inicialmente foram utilizados alguns métodos já conhecidos de processamento de linguagem natural, como *POS tagging* ou a análise semântica das frases para etapas como a extração de conceitos, cuja automatização é uma tarefa mais acessível. Foram também empregados em conjunto com os métodos anteriormente referidos alguns métodos estatísticos de aprendizagem máquina, como, por exemplo, a análise de coocorrência, ainda que tais métodos possam estar a ser utilizados somente para automatizar alguns dos passos do processo de elaboração de ontologias.

Ao longo da década passada foram aparecendo muitos sistemas de elaboração de ontologias de forma (semi)automática utilizando modelos de *Deep Learning*. Este tipo de modelos consegue processar mais informação, de forma mais eficiente e pormenorizada, que os métodos provenientes das abordagens previamente mencionadas, permitindo a obtenção de melhores resultados em diferentes etapas da elaboração da ontologia. Os modelos que são elaborados para automatizar este tipo de tarefas podem não ser modelos convencionais de *Deep Learning*, resultando antes da mistura deste tipo de modelos, tendo sido modificados consoante a(s) etapa(s) do processo que tinham como objetivo automatizar.

1.3 TRABALHO REALIZADO

Nesta dissertação foi descrito, problematizado e fundamentado o processo de idealização e desenvolvimento de uma ontologia para relacionar e caraterizar os conceitos existentes no Livro das Propriedades da Mesa Arcebispal de Braga, ou Tombo da Mitra Arquiepiscopal de Braga (Barros março de 2019), de forma a

proporcionar uma melhor compreensão do seu conteúdo através de uma representação formal do conhecimento nele contido. Este códice, inédito, é constituído por 644 fólios de tamanho grande (19 dedicados ao índice, 622 que constituem o corpo do manuscrito e 3 finais em branco, tendo 2 deles sido posteriormente preenchidos), achando-se em vias de edição (semidiplomática e interpretativa) pela Professora Anabela Barros, do Instituto de Letras e Ciências Humanas da Universidade do Minho, previsivelmente dividido nos quatro livros correspondentes a cada uma das divisões territoriais e religiosas pelas quais se estende a grande maioria das propriedades inventariadas (ou tombadas) e descritas: Comarca de Valença; Comarca de Vila Real; Comarca de Chaves e Comarca de Braga (Braga termo e Braga cidade).

Este códice possui um elevado valor legal e patrimonial, visto que armazena informação bastante concreta das numerosas propriedades pertencentes à Mesa Arcebispal de Braga, no início do século XVII. O seu conteúdo é muito rico, por exemplo, ao nível da terminologia usada para descrever os diferentes tipos de terras e casas, seus proprietários e emprazadores, os produtos nelas cultivados ou espontâneos ou as caraterísticas do terreno e da geografia do norte de Portugal, abrangendo, ainda, propriedades que se estendem até Santarém e à Galiza.

Para o desenvolvimento desta ontologia foram trabalhados os textos já editados por Anabela Barros a partir do manuscrito, na versão semidiplomática, e, sempre que necessário, com a edição interpretativa, já com atualização gráfica, para mais fácil acesso e pesquisa do conteúdo. Em termos gerais, pretendeu-se que a ontologia conseguisse explicitar e relacionar os diferentes conceitos e suas relações, observáveis nos fólios do referido códice. Para que isso fosse possível, foi necessário fazer a implementação de um dicionário com os termos mais comuns do Tombo da Mitra, para se obter uma relação bem fundamentada acerca dos temas mais abordados no códice, e, por conseguinte, os conceitos mais relevantes. Todos estes elementos transformaram-se, mais tarde, nas diversas classes ou propriedades da ontologia. Adicionalmente, pretendeu-se também conceber e desenvolver uma ferramenta de mineração de textos não estruturados para obter especificamente os elementos mais relevantes do Tombo da Mitra, que o permitisse fazer de uma forma automática, reduzindo o seu tempo de análise e permitindo uma caracterização mais detalhada de todos esses elementos. Após a realização desta tarefa, foi necessário integrar na ontologia, também de uma forma automática, a informação obtida através da ferramenta.

Tal como referido por Asim et al. (2018), A aplicação de processos de aprendizagem automática na construção da ontologia pretendida teve grande utilidade, não só pelos erros que evitou no processo de criação da ontologia como também pela redução significativa do tempo que se investiu em tudo aquilo que pudesse ser realizado de forma (semi)automática.

A conceção automática de uma ontologia, no entanto, significa apenas que o seu processo de criação requer menos intervenção humana, e não que a sua elaboração seja totalmente automática. De momento, não se conhece um processo de criação de ontologias completamente automático, não se podendo assim afirmar com certeza que alguma vez isso possa acontecer (Al-Aswadi, Chan e Gan 2020).

1.4 ESTRUTURA DA DISSERTAÇÃO

Para além da introdução, esta dissertação inclui cinco outros capítulos, nomeadamente:

- **Capítulo 2** – Neste capítulo são discutidas a definição, os propósitos e as aplicações de uma ontologia. Também é indicada uma estrutura básica de uma ontologia, acompanhada de um exemplo, com cada um dos seus componentes explicitado. Além disso, este capítulo também revela o modo de elaborar manualmente uma ontologia, seguindo detalhadamente todos os passos de uma metodologia existente. No final da aplicação da metodologia, são enunciados e explicitados alguns problemas derivados da elaboração de ontologias de forma manual.

- **Capítulo 3** – Neste capítulo descreve-se o processo de elaboração de ontologias de forma (semi)automática. Inicialmente são exploradas abordagens linguísticas e de aprendizagem máquina mais antigas, bem como sistemas e modelos criados que utilizam métodos dessas abordagens. Depois, são observadas abordagens mais recentes, envolvendo modelos de *Deep Learning*, fornecendo exemplos da aplicação destes modelos no processamento de texto, que podem ser aplicados à realização de forma automática de etapas específicas de elaboração de uma ontologia.

- **Capítulo 4** – Este capítulo apresenta uma explicação da componente prática desta dissertação, bem como uma avaliação dos resultados práticos obtidos.

- **Capítulo 5** – Este capítulo apresenta a ontologia final obtida, bem como alguns exemplos de interações entre as suas instâncias.

- **Capítulo 6** – Neste último capítulo apresentam-se as conclusões deste trabalho de dissertação, referindo o novo conhecimento obtido em relação aos processos de construção de ontologias de forma (semi)automática, comenta-se a veracidade da ontologia obtida e possíveis alterações que possam vir a ser efetuadas no futuro.

2

ONTOLOGIAS – ESTRUTURA, CONCEITOS E GERAÇÃO MANUAL

2.1 O QUE SÃO ONTOLOGIAS?

Não existe uma definição única ou consensual de *ontologia*. Por exemplo, *Merriam-Webster'sOntology* define *ontologia* tendo em conta uma abordagem mais filosófica: *"a branch of metaphysics concerned with the nature and relations of being or a particular theory about the nature of being or the kinds of existent"*. Porém, mesmo dentro da abordagem informática, existem muitas outras definições, todas, no entanto, seguindo o mesmo molde. Por exemplo, Jiang e Tan (2005) definiram *ontologia* da seguinte forma *"An ontology is an explicit specification of a conceptualization . . . , comprising a formal description of concepts, relations between concepts, and axioms about a target domain"*. Por outro lado Kiong, Palaniappan e Yahaya (2011) definiram-na de forma ligeiramente diferente: *"An ontology can be viewed as a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and relationships between them"*. Ambas as definições são válidas, mas distinguem-se em alguns pequenos aspetos, nomeadamente nos componentes que acham necessário referir enquanto descrevem uma ontologia – Jiang e Tan (2005) mencionam axiomas e regras e excluem atributos de conceitos, enquanto Kiong, Palaniappan e Yahaya (2011) tem o raciocínio inverso. Estas ligeiras diferenças tornam clara a necessidade de encontrar uma definição consensual deste termo. Para se chegar a esta definição é necessário realizar um processo de generalização. Isto é, definir *ontologia* de forma a conseguir que todas as outras definições se possam rever nessa definição.

Uma definição deste género pode ser encontrada em Gruber (1995). Segundo Kim, Caralt e Hilliard (2007), esta definição tem sido referenciada de forma frequente no campo dos sistemas de informação. Nesse artigo, Gruber (1995) afirma o seguinte *"An ontology is an explicit specification of a conceptualization"*. Esta é uma definição bastante genérica e capaz de abranger todas as outras definições, mas, ao mesmo tempo, é sufcientemente concisa para descrever de forma correta uma ontologia. Por conseguinte, é uma boa definição de *ontologia*.

2.2 CONSTITUINTES DE UMA ONTOLOGIA

Estando encontrada uma definição adequada para o conceito de *ontologia*, é necessário explicitar agora a sua estrutura ou constituição. Segundo Al-Aswadi, Chan e Gan (2020) e Zouaq (2011), a estrutura de uma ontologia pode ser definida da seguinte maneira:

$$O =< C, H, R, A >$$

Nesta estrutura, O representa a ontologia em si e os elementos do tuplo as suas diferentes componentes, nomeadamente: C representa o conjunto de conceitos do domínio abordado que são representados na ontologia, H representa a hierarquia de conceitos da ontologia, ou seja as relações taxonómicas existentes, R representa as relações não taxonómicas entre conceitos da ontologia, e, por último, A representa o conjunto de regras ou axiomas definidos na ontologia, a que as instâncias têm de respeitar. De seguida, é descrito cada um dos elementos constituintes referidos anteriormente.

Primeiramente, é necessário definir o que são conceitos. Numa abordagem filosófica (aristotélica), "..., um conceito é definido de acordo com a sua essência, ... conceitos são estruturados de acordo com a diferença específica onde a propriedade essencial é usada para distinguir um conceito de todos os outros conceitos..." (Santos 2010), ou seja, um conceito de um domínio é algo que possui uma essência própria dentro desse mesmo domínio, e não uma caraterística, facto ou ligação entre entidades.

O elemento H do tuplo é representativo das relações taxonómicas existentes entre conceitos. Porém, para que seja possível definir o que são relações taxonómicas, é necessário primeiro definir *taxonomia*. Segundo Rees (2003), *"A taxonomy can thus best be described as a hierarchy created according to data internal to the items in that hierarchy"*, ou seja, uma taxonomia é uma hierarquia de conceitos ou de classes pertencentes a um mesmo domínio. Rees (2003) apresenta uma distinção entre *taxonomia* e *classificação*, indicando que a taxonomia classifica de forma estrutural, de acordo com relações entre diferentes entidades do domínio, enquanto que uma classificação não exige uma hierarquia. O mesmo autor afirma ainda que muitas vezes uma ontologia contém uma hierarquia taxonómica baseada em subclasses. Se uma taxonomia é uma hierarquia de conceitos, então é válido questionar como é que esta hierarquia é construída, ou seja, o que faz com que dois conceitos estejam conectados nesta. A resposta é que estes possuem uma relação direta na hierarquia, se possuírem relações taxonómicas entre eles. Serra, Girardi e Novais (2014) afirmaram que as relações taxonómicas definem uma hierarquia de classes, dando o exemplo de que, se duas classes possuem uma relação taxonómica uma delas é subclasse da outra. Segundo Vazifedoost, Oroumchian e Rahgozar (2007), este tipo de relação é normalmente descoberto nas fases iniciais de conceção da ontologia.

Nem todas as relações estabelecem ligações hierárquicas entre conceitos, pois estes podem estar relacionados por outros motivos. Estes outros tipos de relações são normalmente definidos como relações não taxonómicas. Segundo Vazifedoost, Oroumchian e Rahgozar (2007), as relações não taxonómicas são normalmente encontradas a seguir às relações taxonómicas e, habitualmente, descrevem relações de causalidade, posse ou similaridade. Segundo Serra, Girardi e Novais (2014) existem diferentes tipos de relações não taxonómicas, dependendo de como estas se enquadram no domínio da ontologia. Podem existir relações não

taxonómicas dependentes ou independentes do domínio. As relações não taxonómicas dependentes do domínio possuem esta designação devido ao verbo que intermedia esta relação. Neste tipo de relações o verbo tem de ser caraterístico do domínio. As relações não taxonómicas independentes do domínio correspondem às relações intermediadas por verbos não caraterísticos do mesmo domínio. Estes podem dividir-se em duas subcategorias: agregação e posse. As relações não taxonómicas de agregação descrevem relações parte-todo, ou seja, de meronímia/holonímia. Estas relações não podem ser confundidas com relações de hiponímia ou hiperonímia, visto que estas são, normalmente, caraterísticas de relações taxonómicas. Já as relações não taxonómicas de posse, como o próprio nome indica, descrevem relações nas quais um dos conceitos é "dono" do outro. Serra, Girardi e Novais (2014) descrevem como, na língua portuguesa, as relações não taxonómicas independentes do domínio são representados pela forma possessiva do verbo *ter*, ou pelo uso da preposição *de*, que pode ser, ou não, acompanhada de um artigo definido ou indefinido. As relações não taxonómicas independentes do domínio são sempre representados por estes termos. Porém, o contrário não pode ser garantido, pois estes termos podem ocorrer em frases sem estarem a representar este tipo de relações (Serra, Girardi e Novais 2014).

Por fim, temos o último elemento constituinte do tuplo representativo de uma ontologia: os axiomas, que também são conhecidos como regras. Os axiomas são as afirmações que são consideradas verdadeiras de forma indiscutível no domínio da ontologia, sendo que todas as instâncias de classes criadas para uma ontologia têm de cumprir os axiomas que lhes dizem respeito. A *Merriam-Webster'sAxiom* define *axiom* como *"a statement accepted as true as the basis for argument or inference"*. De acordo com Serra, Girardi e Novais (2014), os axiomas permitem verificar a consistência de uma ontologia e inferir novo conhecimento.

2.3 UM EXEMPLO DE UMA ONTOLOGIA

Após a explicitação dos componentes existentes no tuplo representativo de uma ontologia indicado por Al-Aswadi, Chan e Gan (2020) e Zouaq (2011), vamos agora apresentar um pequeno exemplo de uma ontologia, explicando-a e indicando os seus vários constituintes. O caso de aplicação que escolhemos inclui uma pequena ontologia representativa dos principais intervenientes no mundo musical, e em particular, na gravação de um álbum. Assim, de seguida, será apresentada a forma como esta foi elaborada e, principalmente, como foram obtidos cada um dos componentes mencionados anteriormente, pela ordem previamente estabelecida. Analisemos então este caso:

Foi pedido a um grupo de estudantes de engenharia informática que elaborasse uma aplicação que indicasse alguns dos intervenientes na realização de um álbum e categorização do grupo musical responsável por este. Os elementos do grupo começaram por determinar que um álbum é criado por atos musicais que tanto podem ser grupos musicais, ou artistas a solo. Diferentes artistas podem ter álbuns a solo e ser elementos constituintes de um grupo musical. Para distinguir dos álbuns a solo, os álbuns que um artista ajuda a criar como elemento constituinte de um grupo musical são somente considerados álbuns desse grupo. Atos musicais são normalmente categorizados consoante o género de música que estes tocam. Visto que o universo musical se encontra em constante alteração, existem géneros cuja existência deriva de um outro género. Nestes casos, este género é considerado o supragénero deste conjunto de géneros, que, por sua vez, são designados por

subgéneros. Cada subgénero só pode possuir um supragénero, pois quando um género resulta da mistura de dois ou mais outros géneros, este género designa-se de género de fusão. Por fim, é necessário mencionar que um álbum, para ser lançado, precisa de ser gravado por uma ou mais empresas de gravação e produzido por um ou mais produtores.

Após esta breve descrição do caso que queremos contextualizar, identificaremos o conteúdo de cada um dos componentes desta ontologia. Apresentamos agora os resultados obtidos. Mais à frente, na secção 2.5, este processo será descrito de forma detalhada. Vamos começar o nosso processo pelos conceitos existentes, nomeadamente: Álbum, Ato Musical, Artista, Grupo musical, Género, Empresa de Gravação, Produtor. Depois de termos estabelecido os conceitos representativos das classes da ontologia é necessário definir uma hierarquia de conceitos, com o intuito de descobrir as suas relações taxonómicas. O caso mais óbvio de hierarquia envolve o conceito de *Ato Musical*. Este possui duas relações taxonómicas com os conceitos *Artista* e *Grupo*, uma vez que instâncias destes conceitos podem ser considerados Atos Musicais.

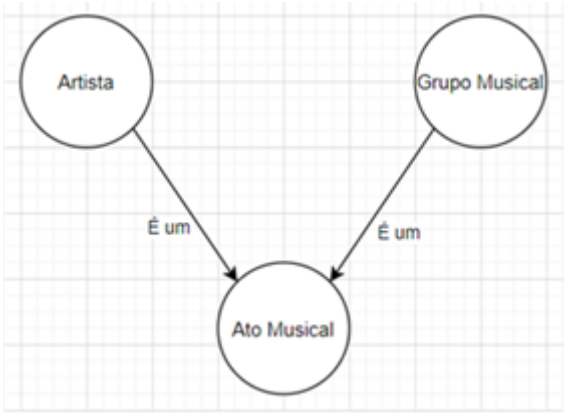

Figura 1: Relação hierárquica entre a classe Ato Musical e as subclasses Artista e Grupo Musical

Outros dois casos representativos de uma hierarquia envolvem o conceito de Género. Pelo texto apresentado para o domínio foi considerado que cada género possui um género do qual deriva e que outros géneros derivam dele, relação essa que se encontra representada na Figura 2. Alguns géneros resultam da junção de dois ou mais géneros. Estes denominam-se *géneros de fusão* e a sua relação com outros géneros encontra-se representada na Figura 3.

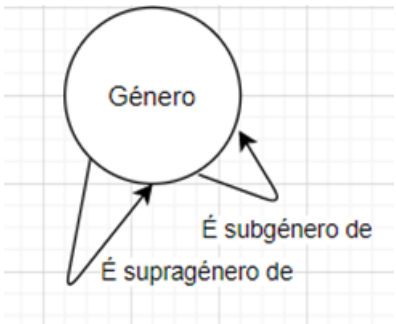

Figura 2: Relação hierárquica entre supragénero e subgénero

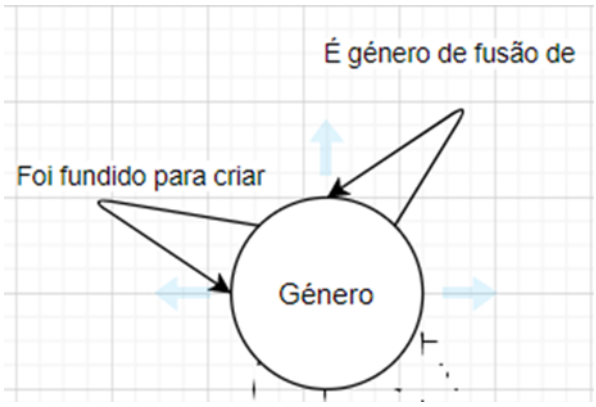

Figura 3: Relação hierárquica entre um género e os géneros cuja fusão o originou

Após a recolha das relações taxonómicas da ontologia é necessário prosseguir para a recolha das relações não taxonómicas. As relações não taxonómicas entre conceitos neste domínio, excluindo uma (apresentada posteriormente), são realizadas através de verbos do domínio e são representadas pelos seguintes verbos: *Gravar, Produzir, Tocar, Criar*. As relações não taxonómicas resultantes destes verbos de domínio são as seguintes:

- Ato Musical criou um Álbum.
- Álbum foi gravado por um Ato Musical.
- Empresa de Gravação gravou um Álbum.
- Álbum foi gravado por uma Empresa de Gravação.
- Produtor produziu um Álbum.
- Álbum foi produzido por um Produtor.
- Ato Musical toca um género.
- Género é tocado por um Ato Musical.

Como foi referido anteriormente, existe um caso de uma relação não taxonómica independente do domínio, nomeadamente a relação entre Artista e Grupo Musical. Esta representa uma relação de agregação (relação parte-todo), uma vez que o Artista é um componente do Grupo Musical.

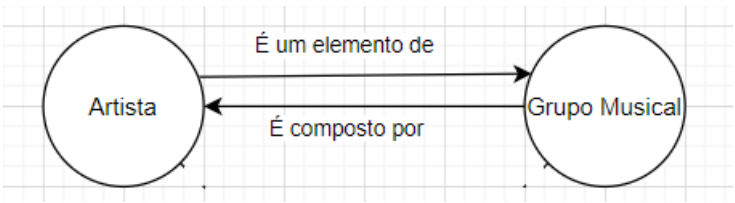

Figura 4: Relação não taxonómica independente do domínio entre as classes Artista e Grupo Musical

Após termos obtido todas as relações definidas, resta fazer a definição dos axiomas que devem ser respeitados por todas as instâncias da ontologia. Os axiomas que foram encontrados são os seguintes:

1. Cada Género pode possuir apenas um Supragénero.
2. Se um artista pertence a um Grupo que lança um Álbum, este não conta como Álbum do Artista.

Concluída a recolha de informação relativa a todos os componentes representados pelo tuplo referido anteriormente, podemos agora fazer a sua representação (Figura 5). Neste esboço da ontologia não explicitámos as regras/axiomas.

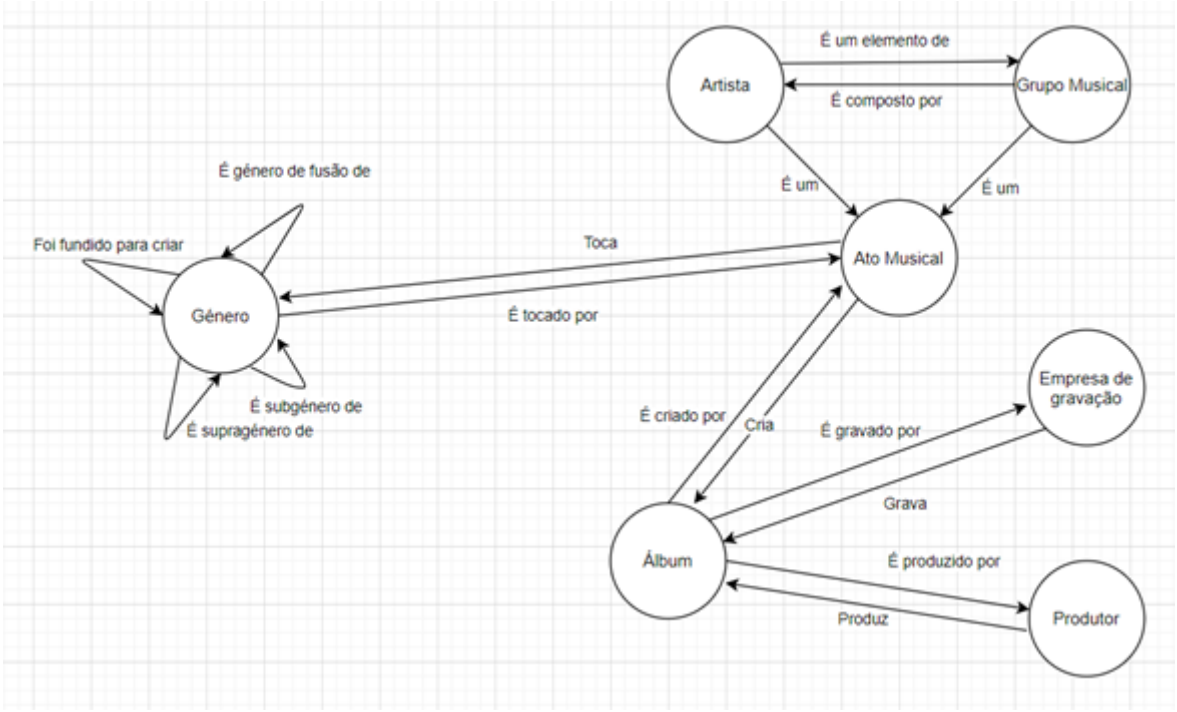

Figura 5: Esquema geral da ontologia criada

O exemplo de ontologia agora apresentado baseia-se no tuplo proposto por Al-Aswadi, Chan e Gan (2020) e Zouaq (2011). No entanto, Serra, Girardi e Novais (2014) e Girardi (2010) apresentam outra proposta de estruturação de uma ontologia, representada pelo tuplo:

$$O = (C, H, I, R, P, A)$$

Este tuplo é similar ao abordado anteriormente, tendo somente as adições dos conjuntos I e P. O conjunto I representa as relações entre as classes ou conceitos e as suas instâncias, enquanto o conjunto P representa as propriedades das classes (Serra, Girardi e Novais 2014), ou seja, esta definição inclui instâncias e atributos de classes como parte integral da ontologia.

2.4 APLICAÇÕES DE ONTOLOGIAS

As ontologias podem ser encontradas em muitos domínios de aplicação. Por exemplo, na área da biologia, Kim, Caralt e Hilliard (2007) afirmaram que *"Recently biological data have increased explosively with an arrival of a new biological research area and new technologies. . . . Making this enormous amount of knowledge sharable and reusable is complex and difficult. Therefore, it becomes a challenge to biology researchers. To address this issue, researchers have turned to ontology"*, ou seja, alertaram para o quão importantes são as ontologias para a estruturação de quantidades grandes de informação, para que esta possa ser consultada de forma eficiente, especialmente se a informação relativa a essa área estiver a crescer de forma quantitativa, a um ritmo elevado.

Relativamente à área da saúde, Kiong, Palaniappan e Yahaya (2011) afirmaram que *"Most healthcare institutions such as hospitals and clinics store their data in the form of databases of various formats. The Health Ontology System that we have developed provides a means to integrate these data with concepts and semantics in the form of a shared cumulative ontology for enabling machines to interpret them"*, ou seja, referiram como, no sistema por eles desenvolvido, se utilizou a estrutura de uma ontologia para ajudar a automatizar processos em instituições de saúde. Quanto à estrutura de procedimentos negociais, Fan, Zhang e Sun (2009) afirmaram que *"... semantic gaps exist among organizations when integrating their business processes. Different companies use different terminology even when denoting an identical concept. These days, ontology theory is adopted widely to resolve the problems of semantics, e.g., enterprise information integration and semantics search engine. We apply ontology and ontology related approaches to business process integration"*, ou seja, neste caso o uso de ontologias é necessário para interligar termos diferentes que podem ser usados para representar a mesma informação, prevenindo que existam falhas de comunicação e tornando este tipo de procedimentos mais eficiente.

Como pudemos constatar, a aplicação de ontologias não só é necessária, como, em certos casos, se revela essencial para o bom funcionamento de instituições ou para a existência de progresso em algumas áreas. Song et al. (2014) afirmaram que *"Now ontology... plays a prominent role in many fields"*. Olhando para os exemplos referidos anteriormente, esta afirmação assume-se absolutamente verdadeira. Desta forma terminamos a explicação base de uma ontologia, em termos de estrutura, constituição e aplicações. Na secção seguinte abordaremos com maior pormenor o processo de elaboração ontológico de forma manual, apresentando de seguida as suas desvantagens.

2.5 ELABORAÇÃO MANUAL DE UMA ONTOLOGIA

2.5.1 *Metodologia para a construção de uma ontologia*

O processo de construção de uma ontologia de forma manual é, como outros processos elaborados totalmente pelo ser humano, altamente subjetivo. Esta subjetividade culmina na existência de abordagens ligeiramente distintas, ainda que estas possuam uma base comum. Nesta dissertação utilizámos a metodologia descrita em Noy e McGuinness (2001). Segundo estes autores, primeiramente, começa-se por identificar o domínio de trabalho e verificar se sobre ele já existem ontologias realizadas. De seguida, é necessário decidir

quais os termos mais relevantes do domínio, a partir dos quais se obtêm as entidades (classes), as relações entre as entidades (*object properties*) e os atributos das entidades (*data properties*). Por fim, realiza-se o povoamento da ontologia.

Segundo estes autores, podem-se retirar conclusões acerca de qual o domínio da ontologia ao responder a quatro questões, nomeadamente:

1. Qual o domínio que a ontologia cobrirá?
2. Qual será o uso da ontologia?
3. A que tipo de questões deverá a informação da ontologia proporcionar respostas?
4. Quem irá utilizar e efetuar a manutenção da ontologia?

Noy e McGuinness (2001) afirmaram que, apesar de as respostas a estas questões poderem ser alteradas à medida que o processo de desenho de ontologia prossegue, estas podem sempre ajudar a restringir a abrangência do modelo.

Após a determinação do domínio representado pela ontologia, é necessário considerar a utilização de ontologias já existentes e refiná-las de modo a representarem o domínio de forma adequada. Estes autores chamaram a atenção para o facto de que o uso de ontologias já existentes pode ser um imperativo, caso o sistema a desenvolver necessite de interagir com uma aplicação que já se tenha comprometido com uma ontologia ou um conjunto de vocábulos controlado. Além disso, estes autores também fizeram referência a bibliotecas de ontologias existentes, como a Ontolíngua e a DAML. Se após este passo se chegar à conclusão de que ainda não existe uma ontologia que explicite o que se pretende conhecer acerca do domínio abordado, então é necessário prosseguir com este processo. Se existir uma ontologia que descreva de forma adequada o domínio que queremos conceptualizar, então o processo é concluído neste passo. Caso não exista uma ontologia que descreva o domínio de uma forma considerada adequada, o seu processo de elaboração prossegue para o passo seguinte, que consiste na enumeração de termos do domínio que sejam relevantes para o mesmo. Este é um passo bastante relevante da ontologia, visto que os termos recolhidos provavelmente farão parte da constituição da ontologia. Segundo Noy e McGuinness (2001), inicialmente a lista de termos que é necessário elaborar pode conter vocábulos que representem o mesmo conceito, que tenham as mesmas propriedades, ou relativamente aos quais haja dúvidas quanto à sua natureza de classes ou atributos.

Após o término da enumeração de todos os termos considerados relevantes na ontologia, são selecionados desta lista as classes ou conceitos da ontologia, procurando relações taxonómicas entre eles, com o intuito de construir a hierarquia de classes. Noy e McGuinness (2001) fizeram referência a abordagens para este passo que sejam *top-down*, *bottom-up* ou uma combinação das duas, afirmando que o método que um criador da ontologia usa depende da visão pessoal que este tem do domínio. Uma abordagem *top-down* implica a determinação dos conceitos mais genéricos, sendo que depois é efetuado um processo de especialização para descobrir conceitos mais específicos. Uma abordagem *bottom-up* é a abordagem oposta da anterior, pois esta implica primeiramente determinar os conceitos mais específicos do domínio, sendo depois aplicado a estes um processo de generalização para determinar outros conceitos. Por fim, a última abordagem que

referimos resulta da combinação das duas anteriores, ou seja, primeiro descobre-se um conjunto de conceitos e depois realizam-se processos de generalização e especialização para determinar os outros conceitos existentes. Independentemente da abordagem que se utiliza, é normal começar com a determinação das classes, que habitualmente são representadas por termos independentes que não descrevem outros objetos. A hierarquia entre classes é obtida quando questionamos se o facto de um elemento pertencer a uma classe faz com que esse elemento pertença automaticamente a outra.

Após a definição de classes ter sido completada, é necessário descrever a estrutura interna dos conceitos. Segundo Noy e McGuinness (2001), é provável que a maior parte dos termos que sobram do passo de enumeração de termos importantes seja propriedade das classes já descobertas. Por conseguinte, são determinadas as classes que estas propriedades caraterizam. Cada subclasse herda as propriedades caraterísticas da classe que se encontra acima na hierarquia. Consequentemente, as propriedades devem ser atribuídas à classe mais geral (ou mais acima na hierarquia) que as possui. Segundo estes autores, existem alguns tipos de propriedades que se podem tornar em atributos de uma ontologia. Estas são as propriedades intrínsecas a uma classe, as propriedades extrínsecas a uma classe, as partes constituintes de um objeto, caso este possua uma estrutura (podem ser partes físicas ou abstratas), e as relações, que tanto podem ser estabelecidas com outras classes como com elementos pertencentes à mesma classe.

Após a conclusão da definição dos atributos, é necessário caraterizá-los. Noy e McGuinness (2001) indicaram que atributos diferentes podem possuir caraterísticas diferentes e fizeram referência a algumas delas, como a cardinalidade, o tipo de valor, o domínio e o alcance. A cardinalidade indica quantos valores distintos um atributo pode ter, o tipo de valor indica quais os valores que o atributo pode assumir, (isto é, se os valores representativos deste são boleanos, números, entre outros), e o domínio e o alcance representam, respetivamente, as classes a que a propriedade é atribuída ou que de alguma forma descreve, e as classes afetadas pela propriedade de uma outra classe do domínio (como é o caso nas relações).

De seguida, referimos aquele que é o último passo da elaboração de uma ontologia, a criação de instâncias das suas classes. Segundo Noy e McGuinness (2001), a criação de uma instância individual de uma classe obriga ao seguimento de um conjunto de passos. Primeiramente, é escolhida a classe a representar na instância, algo a que se segue a criação de uma instância individual dessa mesma classe. O último passo deste processo consiste no preenchimento dos valores dos atributos das instâncias.

2.5.2 *Elaboração da ontologia musical*

Para comprovar a validade da metodologia apresentada anteriormente, na secção 2.3, vamos aplicá-la a um caso concreto. Para tal, reutilizaremos a ontologia musical, seguindo de forma detalhada a metodologia proposta por Noy e McGuinness (2001). Assim, reescrevemos o texto representativo da ontologia para que seja possível incorporar propriedades intrínsecas e extrínsecas aos conceitos, o que, juntamente com o último passo do processo de criação de ontologias de Noy e McGuinness (2001), faz com que o tuplo representativo da ontologia seja mais similar ao de Serra, Girardi e Novais (2014) e Girardi (2010), referido anteriormente. Para auxiliar na construção da ontologia utilizaremos o sistema *Protégé*, uma plataforma de investigação e

desenvolvimento de sistemas baseados em conhecimento. O funcionamento deste sistema pode ser consultado em Gennari et al. (2003). O texto reescrito no qual nos baseámos para a construção da ontologia é o seguinte:

Foi pedido a um grupo de estudantes de engenharia informática que elaborasse uma aplicação, que indicasse alguns dos intervenientes na realização de um álbum e categorização do grupo musical responsável por este.

Os elementos do grupo começaram por determinar que um álbum é criado por atos musicais que são ou grupos musicais, ou artistas a solo. Diferentes artistas podem ter álbuns a solo e ser elementos constituintes de um grupo musical. Para distinguir dos álbuns a solo, os álbuns que um artista ajuda a criar como elemento constituinte de um grupo musical, são somente considerados álbuns desse grupo.

Um álbum é constituído por um identificador único, um nome, uma breve descrição, uma data de lançamento, uma lista de identificadores de quem o criou, uma lista de identificadores de quem o produziu e uma lista de identificadores das empresas de gravação onde este foi gravado.

Um artista é composto por um identificador único, uma data de nascimento, uma data de falecimento (se existir), o local de nascimento, o nome que lhe foi dado à nascença, um nome pelo qual é conhecido, uma breve descrição, uma data a indicar o início da atividade, uma data a indicar o fim da atividade (se existir), uma lista de géneros que o artista toca ou tocou no passado, uma lista de álbuns já criados pelo artista e a lista de grupos a que este pertence ou pertenceu.

Um grupo é composto por um identificador único, o local de origem, um nome pelo qual é conhecido, uma breve descrição, uma data a indicar o início da atividade, uma data a indicar o fim da atividade (se existir), uma lista de géneros que o artista toca ou tocou no passado, uma lista de álbuns já criados pelo artista e a lista de artistas que já participaram nele ou que ainda participam.

Visto que o universo musical se encontra em constante alteração, existem géneros cuja existência deriva de um outro género. Este género é considerado o supragénero deste conjunto de géneros que derivam deste (seu predecessor), designados de subgéneros. Cada subgénero só pode possuir um supragénero, pois quando um género resulta da mistura de dois ou mais outros géneros este designa-se de género de fusão.

Um género é composto por um identificador único, um nome, uma pequena descrição, o seu supragénero (se existir), uma lista de subgéneros (se existirem), uma lista de géneros cuja fusão originou o próprio género (se existirem), uma lista de géneros que resultaram da fusão deste género com outro e uma lista de atos musicais que tocam ou tocavam o género.

Por fim, é necessário mencionar que um álbum para ser lançado precisa de ser gravado por uma ou mais empresas de gravação e produzido por um ou mais produtores.

Um produtor é composto por um identificador único, um nome pelo qual é conhecido, um nome dado à nascença, uma data de nascimento, uma data de falecimento (se existir), uma data de início de atividade, uma data de fim de atividade (se existir) e uma lista de álbuns produzidos.

Uma empresa de gravação é composta por um identificador único, um nome, o nome do CEO desta, o ano de fundação, a sua sede e uma lista de álbuns gravados e distribuídos por esta.

Iniciaremos agora a aplicação da metodologia elaborada por Noy e McGuinness (2001), começando pela definição do domínio. O passo referente à definição do domínio, está concluído aquando da apresentação das seguintes respostas às perguntas referidas anteriormente para este mesmo passo:

1. Qual o domínio que a ontologia cobrirá?

   O domínio que a ontologia cobrirá é relativo ao processo de elaboração e lançamento de um álbum.

2. Qual será o uso da ontologia?

   Esta ontologia será desenvolvida com um intuito de fazer uma demonstração do processo referido de elaboração de ontologias.

3. A que tipo de questões deverá a informação da ontologia proporcionar respostas?

   A ontologia elaborada deverá responder a quaisquer perguntas que envolvam somente os conceitos encontrados e suas propriedades.

4. Quem irá utilizar e efetuar a manutenção da ontologia?

   Não será necessário efetuar qualquer tipo de manutenção em relação a esta ontologia, visto que esta só foi elaborada com o propósito demonstrativo referido anteriormente, não havendo seguimento posterior à sua elaboração após esta etapa da dissertação.

Concluído o processo de resposta a estas quatro questões temos a definição do domínio da ontologia validada. Assim, podemos prosseguir para o próximo passo, que diz respeito à ponderação do uso de outras ontologias já existentes para este domínio. Em relação a este passo do processo de desenvolvimento ontológico, para podermos realizar os passos seguintes, consideraremos que não existem ontologias que melhor representem este domínio. Se o propósito principal desta ontologia não fosse a realização de uma demonstração, esta poderia não ser a abordagem tomada, visto que o trabalho requerido para a elaboração dos próximos passos poderia ser evitado sem que existisse perda de qualidade na ontologia obtida no final.

Com a decisão de não utilizar uma ontologia já existente e de, consequentemente, continuar a aplicação desta metodologia, a elaboração da nossa ontologia prossegue com a enumeração de termos considerados relevantes para o domínio.

Para podermos determinar e enumerar todos os termos relevantes para este domínio, é necessário reler com atenção o texto anterior a fim de verificar quais são os constituintes relevantes do domínio. Após uma leitura atenta, os termos que obtivemos foram os seguintes: Álbum, Criado por, Atos musicais, São, Grupos, Artistas, Ser elementos constituintes de, Identificador único, Nome, Descrição, Data de lançamento, Lista de atos, Lista de produtores, Lista de empresas de gravação, Início de atividade, Fim de atividade, Lista de géneros, Toca/tocou, Lista de álbuns, Data de nascimento, Data de falecimento, Local de nascimento, Nome dado à nascença, Lista de grupos, Pertence/pertenceu,Local de onde originou, Lista de artistas, Participam/participaram, Deriva de, Precede, Supragénero, Subgéneros, Género de fusão, Lista de subgéneros, Fusão originou, Resultam da fusão de, Distribuído por, Empresa de gravação, Produzido por, Produtor, CEO, Ano de fundação, Sede.

Estando concluída a enumeração dos termos, passamos ao passo da definição de classes ou de conceitos e à descoberta da hierarquia que estes estabelecem. A abordagem que escolhemos para efetuar este passo foi uma abordagem combinada. Uma primeira abordagem levou a que chegássemos a um conjunto de conceitos bastante concretos, visto que estes não estabelecem relações com outros conceitos e não são caraterísticos de nenhum conceito. Esse conjunto inclui os conceitos: Álbum, Grupo, Artista, Género, Empresa de Gravação e Produtor.

Estando recolhidos os primeiros conceitos, foi necessário efetuar processos de generalização e especialização. Ao generalizar podemos verificar que quer um grupo musical quer o artista são atos musicais, o que nos permite concluir que são suas subclasses. Além disso, também podemos determinar que existem termos que são listas de conceitos (i.e. lista de álbuns, lista de géneros, entre outros). Porém, estas listas não são reconhecidas como conceitos, visto que elas representam a cardinalidade de relações/*object properties* futuras com outros conceitos. Logo ao fim desta etapa de generalização a lista de conceitos que obtivemos foi a seguinte: Álbum, Grupo, Artista, Género, Empresa de gravação, Produtor, Ato Musical.

Ao efetuarmos o processo de especialização podemos ter algumas dúvidas acerca do que considerar como conceitos mais específicos, quando falamos do conceito de género. Porém, apesar de poderem parecer conceitos próprios, o supragénero e os subgéneros não são conceitos, mas sim relações entre instâncias do mesmo conceito, o género. Esta conclusão é de fácil compreensão e resulta do facto de não haver nenhuma caraterística diferenciadora desses conceitos para com o género. Logo ao fim desta etapa de especialização a lista de conceitos obtida foi a seguinte: Álbum, Grupo, Artista, Género, Empresa de gravação, Produtor, Ato Musical. A lista obtida no final deste processo de especialização é, por conseguinte, a lista de conceitos completa, sendo as classes Grupo e Artista subclasses de Ato Musical e constituindo estas as únicas relações hierárquicas entre classes, ou seja, as únicas relações taxonómicas desta ontologia.

De seguida, encontra-se apresentada a estrutura hierárquica das classes da nossa ontologia, através do sistema *Protégé*, que foram obtidas com a execução do último passo. Com a conclusão deste passo, o processo prossegue com a identificação dos atributos relativos a estas classes.

Figura 6: Hierarquia de classes da ontologia musical

Neste passo começaremos por avaliar os termos que sobram da enumeração de termos dividindo-os entre os quatro tipos de propriedades referidos anteriormente. Os dois primeiros tipos de propriedades deverão ser representativos de atributos de conceitos (*data properties*), enquanto o último tipo de propriedade diz respeito a relações entre conceitos (*object properties*), uma distinção necessária e relevante para a elaboração da ontologia

no *Protégé*. O terceiro tipo de propriedade é representado por relações parte-todo. Estas tanto podem resultar em *data properties* como em *object properties*. Enquanto esta separação é realizada, são também aplicadas estas propriedades aos conceitos a que dizem respeito. As primeiras propriedades que abordaremos são as intrínsecas às classes ou conceitos. Estas caraterizam o conceito somente em relação às suas caraterísticas internas.

Após uma avaliação do texto relativo a este domínio e aos termos enumerados foi possível chegar a algumas propriedades intrínsecas, que são referidas de seguida, começando com as que se aplicam ao maior número de classes e prosseguindo com as mais específicas. O identificador único, o nome (pelo qual a instância é conhecida) e a descrição são propriedades intrínsecas de todos os conceitos. É relevante apontar que, por serem propriedades da superclasse Ato Musical, estas são herdadas pelas subclasses Artista e Grupo. A data de nascimento, a data de falecimento e o nome dado à nascença são propriedades intrínsecas aos conceitos de Artista e Produtor. Como são propriedades caraterísticas de uma das suas subclasses, estas não são propriedades do Ato Musical. As datas representativas do início e final de atividade são propriedades intrínsecas caraterísticas dos conceitos de Artista, Grupo e Produtor. O ano de fundação e o CEO são propriedades intrínsecas do conceito de Empresa de Gravação. A data de lançamento é uma propriedade intrínseca do conceito de Álbum.

Tendo concluído a fase de recolha de propriedades intrínsecas, prosseguimos para a recolha de propriedades extrínsecas. Estas caraterizam o conceito somente em relação às suas caraterísticas externas. Após uma avaliação do texto relativo a este domínio e aos termos enumerados foi possível chegar a estas propriedades extrínsecas: o local de nascimento é uma propriedade extrínseca dos conceitos de Artista e Produtor; o local de onde originou é uma propriedade extrínseca do conceito de Grupo; e a lista de fundadores e o local onde se localiza a sede são propriedades extrínsecas do conceito de Empresa de Gravação.

De seguida, na Figura 7 podemos ver todas as *data properties* existentes nesta ontologia descritas no sistema *Protégé*. Estas são representativas dos dois primeiros tipos de propriedades.

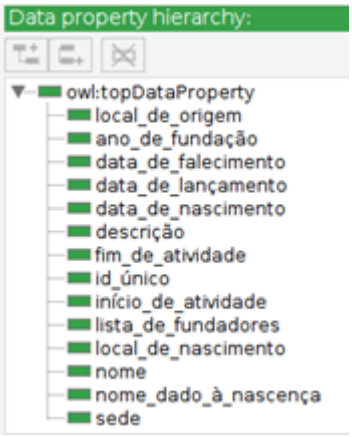

Figura 7: *Data properties* (atributos) existentes na ontologia musical

As propriedades referidas anteriormente dizem respeito a atributos de conceitos, ao contrário das propriedades seguintes, que serão relativas a relações entre conceitos nas circunstâncias desta ontologia. Começaremos

por abordar as propriedades que são representativas de relações parte-todo. As propriedades que cabem dentro desta categoria são representativas de relações de parte-todo ou de elemento-conjunto, ou seja, relações de meronímia/holonímia e hiponímia/hiperonímia, de forma similar às relações não taxonómicas independentes do domínio. No caso da nossa ontologia só existe um exemplo de uma propriedade deste tipo, sendo esta a que descreve a relação hiponímia/hiperonímia entre os conceitos de artista e de grupo, que pode ser observada no termo multipalavra "ser elementos constituintes de" e nos termos "pertence/pertenceu" e "participou/participaram". Neste caso esta propriedade descreveu uma relação entre dois conceitos, porém este caso nem sempre acontece, podendo, por exemplo, o caso apresentado em Noy e McGuinness (2001), de pratos numa refeição, ser interpretado como um conjunto de *data properties* ou uma relação entre um conceito representativo de uma refeição e outro representativo de um prato. Este caso é um exemplo de subjetividade existente na elaboração de ontologias, algo que, como verificaremos, não é um acontecimento particularmente raro.

O quarto e último tipo de propriedade que abordaremos diz respeito às *object properties*, que identificam relações não taxonómicas entre conceitos através de verbos caraterísticos do domínio. Para identificar estas propriedades basta observar os termos que restam. A partir desses termos podemos concluir que: Ato Musical toca/tocou lista de Géneros, Álbum é criado por lista de Atos Musicais, Álbum produzido por lista de Produtores, Álbum distribuído por Empresa de Gravação, Género deriva de Género, Género precede Géneros, Género resulta da fusão de Géneros. Estas propriedades, mais as propriedades relativas às relações inversas, são as propriedades representativas deste tipo. A sua recolha conclui esta fase da construção da ontologia. De seguida, na Figura 8, podemos ver todas as *object properties* existentes nesta ontologia descritas no sistema *Protégé*.

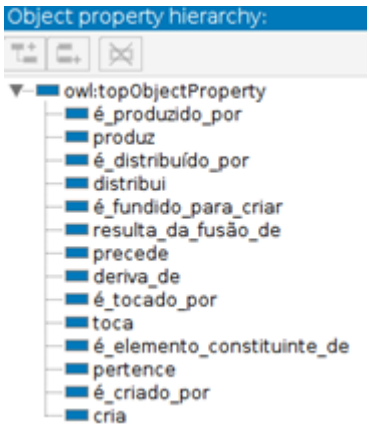

Figura 8: *object properties* (relações) da ontologia musical

Estas propriedades são representativas dos dois últimos tipos de propriedades nestas circunstâncias. A elaboração da ontologia prosseguiu com a caraterização das propriedades obtidas nesta fase.

Neste passo da elaboração da ontologia foi necessário caraterizar as propriedades obtidas indicando algumas das suas caraterísticas definitivas. Nesta etapa abordaremos quatro dessas caraterísticas, que também foram as abordadas por Noy e McGuinness (2001), nomeadamente o tipo de valor, a cardinalidade, o domínio e o alcance. Abordaremos estas quatro caraterísticas seguindo esta ordem.

O tipo de valor define quais os formatos que os valores de uma propriedade podem tomar. Como *object properties* são a representação de relações entre classes, o tipo de valor assume-se como mais relevante para

data properties. De seguida, encontram-se descritos os tipos de valor das diferentes *data properties* obtidas: Identificador único, nome pelo que são conhecidos, descrição, nome dado à nascença, local de nascimento, local de origem de um grupo, CEO de uma empresa de gravação e local onde se localiza a sede da empresa possuem tipo de valor STRING. Data de nascimento, data de falecimento e data de lançamento possuem tipo de valor DATETIME. Ano de fundação, início de atividade e fim de atividade são do tipo INTEGER.

A cardinalidade define quantos valores uma propriedade tem. A cardinalidade das diferentes *data properties* é a seguinte: Identificador único, nome pelo que são conhecidos, descrição, nome dado à nascença, local de nascimento, local de origem de um grupo, local onde se situa a sede da empresa, data de nascimento, data de falecimento, início de atividade, fim de atividade, data de lançamento, ano de fundação, CEO e relação com um supragénero, possuem como cardinalidade um, enquanto todas as *object properties*, exceptuando a relação de um género com o seu supragénero, têm vários valores de cardinalidade.

O domínio de uma *data property* é a classe à qual essa propriedade pertence, enquanto o seu alcance é o conjunto de valores que a propriedade pode ter. As *object properties* possuem como domínio a classe que efetua a *object property* e como alcance a classe sobre a qual a *object property* é efetuada. Relativamente às *data properties* os seus alcances são: xsd:string para identificadores únicos, nome pelo qual é conhecido, descrição, local de origem, CEO, sede, nome dado à nascença e local de nascimento; xsd:integer para ano de fundação, início de atividade e final de atividade; xsd:dateTime para data de lançamento, data de nascimento, data de falecimento. Neste caso, as restrições relativas ao alcance correspondem ao tipo de valor das *data properties*, podendo não ser este o caso, se o valor tivesse que estar contido dentro de um conjunto pré-definido.

Relativamente às *data properties* os seus domínios são: todas as classes para identificadores únicos, nome pelo qual é conhecido e descrição; classes Artista, Grupo e Produtor para o início de atividade e final de atividade; classes Artista e Produtor para a data de nascimento, data de falecimento, nome dado à nascença e local de nascimento; classe Grupo para o local de origem; classe Álbum para a data de lançamento; classe Empresa de Gravação para o ano de fundação, o CEO e a sede.

Tendo já referido os domínios das *data properties*, resta apenas enumerar os domínios e alcances das *object properties*. Uma relação inversa àquela descrita por uma *object property* tem como domínio o seu alcance e como alcance o seu domínio. Os domínios e os alcances das *object properties* estão apresentados na Tabela 1.

Para concluir a elaboração da ontologia musical foi necessário executar o último passo deste processo: a criação das instâncias dos conceitos obtidos. Assim, foi criada pelo menos uma instância para cada conceito, sabendo-se que todas elas estarão relacionadas umas com as outras no final do processo, através das respetivas *object properties*. Toda a informação contida nas instâncias elaboradas será hipotética, não sendo representativa de qualquer cenário real.

A primeira instância a mencionar é a instância 'Álbum1', que, como o nome indica, é uma instância da classe "Álbum". Na Figura 9 podemos observar esta instância criada no sistema *Protégé*.

object property	Domínio	Alcance	Relação inversa
Ato musical cria Álbum	Ato Musical	Álbum	Álbum é criado por Ato Musical
Artista pertence a Grupo	Artista	Grupo	Grupo é constituído por Artistas
Produtor produz Álbum	Produtor	Álbum	Álbum é produzido por Produtor
Empresa de gravação lança Álbum	Empresa de gravação	Álbum	Álbum é lançado por Empresa de Gravação
Género deriva de Género	Género	Género	Género precede Género
Género resulta da fusão de Géneros	Género	Género	Género é fundido para gerar Género
Ato musical toca Género	Ato musical	Género	Género é tocado por Ato Musical

Tabela 1: Domínio e alcance de *object properties*.

Figura 9: Uma instância da classe "Álbum"

A instância que considerámos a seguir foi a instância 'Artista1', que, como o nome indica, é uma instância da classe "Artista". De seguida, na Figura 10, podemos ver a sua descrição no sistema *Protégé*.

Figura 10: Uma instância da classe "Artista"

A próxima instância é a instância 'Grupo1'. Esta é uma instância da classe "Grupo". Na Figura 11 podemos ver a sua descrição no sistema *Protégé*.

Figura 11: Uma instância da classe "Grupo"

Os próximos exemplos são as instâncias 'Género1' (Figura 12) e 'Produtor1' (Figura 13), que são instâncias da classe "Género" e da classe "Produtor", respetivamente. Por fim, temos a instância 'Empresa1' (Figura 14), que constitui uma instância da classe "Empresa de Gravação".

Figura 12: Uma instância da classe "Género"

Figura 13: Uma instância da classe "Produtor"

Figura 14: Instância da classe "Empresa De Gravação"criada

De seguida, para demonstrar um exemplo da caraterização de uma *data property* e da caraterização de uma *object property*, encontram-se descritas as caraterísticas da *data property* Data de Lançamento do "Álbum" e da *object property* "produz" no sistema *Protégé*. A *data property* indica a data em que a instância da classe "Álbum" foi disponibilizada para venda, enquanto a *object property* descreve a relação de produção de um "Álbum" entre as classes "Produtor" e "Álbum".

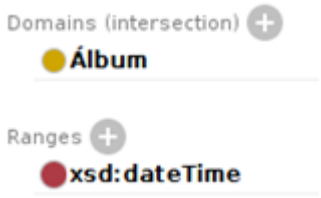

Figura 15: Exemplos de caraterísticas de uma *data property*

Figura 16: Exemplos de algumas caraterísticas de uma *object property*

Com a apresentação destes exemplos, concluímos o processo de elaboração da nossa ontologia musical seguindo a metodologia de Noy e McGuinness (2001).

2.5.3 *Desvantagens da elaboração manual de ontologias*

Como vimos, a ontologia musical que desenvolvemos anteriormente foi elaborada de forma manual. Isso significa que todos os passos e decisões tomadas ao longo do processo de elaboração da ontologia foram realizadas inteiramente pelos seus construtores, sem auxílio de qualquer sistema. Porém este não é o único método de elaboração de ontologias. As ontologias também podem ser elaboradas de forma (semi)automática, ou

seja, de forma que parte do seu processo de construção seja realizado por um sistema e não por um engenheiro de conhecimento. De entre estas duas possibilidades, a elaboração de uma ontologia de forma manual ainda é a maneira mais comum. Porém, esta possibilidade tem algumas consequências negativas. A captação de termos que são relevantes para uma dada ontologia, por exemplo, depende quase inteiramente da intuição de quem estrutura a própria ontologia, especialmente quando o domínio de trabalho envolve uma grande quantidade de textos não estruturados. Tal circunstância não facilita a perceção que o criador da ontologia possa ter acerca de quais são os termos mais comuns, e de como as relações entre os vários elementos constituintes são estabelecidas, algo que é frequentemente muito subjetivo (Al-Aswadi, Chan e Gan 2020).

Outro problema da abordagem manual é o facto de esta requerer imenso tempo para a sua realização. Isto deve-se, essencialmente, ao processo de leitura e de interpretação de uma grande quantidade de textos do domínio a ser abordado, processo esse que necessita de bastante tempo para a sua realização (Al-Aswadi, Chan e Gan 2020).

Em domínios de dimensão relativamente pequena a elaboração manual de uma ontologia já consegue ser problemática. Quando as ontologias passam a ser de grande dimensão, um cenário que se está a tornar cada vez mais comum (como foi possível constatar na secção 1 desta dissertação), as dificuldades resultantes do processo de elaboração manual assumem uma dimensão de tal complexidade que esta se torna numa impossibilidade.

São estes problemas e a sua dimensão que tornam vital o desenvolvimento de sistemas de elaboração de ontologias de forma semiautomática, sendo estes tipos de sistemas necessários para a estruturação dos grandes conjuntos de dados e informação, cuja complexidade continuará, progressivamente, a aumentar.

3

GERAÇÃO AUTOMÁTICA DE ONTOLOGIAS

Nos últimos anos tem havido um aumento de conhecimento sem precedentes, nas mais variadas áreas do saber. O aumento exponencial do conhecimento obtido tem consequências nefastas na construção e verificação de uma ontologia, uma vez que provoca um aumento ainda mais acentuado do tempo dedicado à sua construção, e implica subjetividade da sua construção e erro associado à análise de um domínio de grandes proporções. Este é um problema que tem sido abordado de forma séria ao longo das últimas décadas, originando várias propostas para novas abordagens (semi)automáticas de *Ontology Learning*.

Os sistemas que foram elaborados, seguindo este tipo de abordagens de automatização do processo de construção de ontologias, envolveram durante muitos anos o uso de diferentes técnicas, provenientes de áreas diferentes, como o processamento de linguagem natural ou a aprendizagem máquina. Algumas técnicas destas áreas foram muitas vezes combinadas com o intuito de atingir resultados melhores neste processo. Mais recentemente, numa tentativa de automatizar todo o processo de *Ontology Learning*, foram construídos sistemas que tentam solucionar este problema com base em arquiteturas de *Deep Learning*. Neste momento, a automatização do processo de construção de ontologias ainda não é uma realidade. Apesar de ainda não ser uma realidade o total automatismo deste processo, os passos que se têm dado nesta área têm sido muito importantes, permitindo agilizar de modo considerável a elaboração e a manutenção de ontologias de forma mais objetiva e com limitado *input* do engenheiro de conhecimento.

3.1 ELABORAÇÃO DE UMA ONTOLOGIA DE FORMA (SEMI)AUTOMÁTICA

Anteriormente, referimos que existem vários problemas relacionados com a elaboração de ontologia de forma manual. Numa tentativa de solucionar tais problemas pode-se proceder à elaboração semiautomática ou automática de ontologias, que é, em termos gerais, caraterizada pela construção de uma ontologia que é feita com uma intervenção muito limitada dos utilizadores e dos peritos do domínio do conhecimento em questão.

Uma elaboração semiautomática de uma ontologia implica uma intervenção maior dos utilizadores e dos peritos do que uma elaboração automática. A conceção automática de uma ontologia, no entanto, significa apenas que o seu processo de criação requer menos intervenção humana e não que a sua elaboração seja totalmente automática. De momento, tanto quanto o nosso conhecimento o permite, não se conhece um processo de criação de ontologias completamente automático, não se podendo, assim, afirmar com certeza que alguma

vez isso possa acontecer (Al-Aswadi, Chan e Gan 2020). Nesta parte da dissertação analisaremos diferentes abordagens ao processo de *Ontology Learning*.

O processo de *Ontology Learning* a partir de textos é um processo (semi)automático de construção de ontologias referente à extração de elementos ontológicos de um *input* textual, elaborando uma ontologia a partir deles (Shamsfard 2006). De seguida, observaremos o desenvolvimento deste processo, analisando cada uma das suas etapas.

Buitelaar, Cimiano e Magnini (2004) organizaram um conjunto de etapas, seis na sua totalidade, que designaram por *Ontology Learning Layer Cake*, para o desenvolvimento de uma ontologia, começando com a realização das tarefas menos complexas, que se desenvolvem para a obtenção e extração de conceitos, e acabando com tarefas mais complexas, como a definição de regras que retratam os factos implícitos da ontologia. As etapas constituintes deste processo são, por ordem de execução e complexidade, a extração de termos, a descoberta de sinónimos, a formação de conceitos, a hierarquia de conceitos, a descoberta de relações e, por último, a extração de regras ou axiomas. De seguida, encontram-se descritas as várias camadas do referido "bolo".

A primeira das etapas descritas na metodologia de Buitelaar, Cimiano e Magnini (2004) é a de extração de termos. Esta é um pré-requisito para todos os aspetos de *Ontology Learning* a partir de texto. Estes termos representam elementos específicos do domínio a analisar, sendo, por conseguinte, a sua extração essencial. A segunda etapa é relativa à descoberta de sinónimos. Num conjunto de textos existem muitos termos que possuem o mesmo significado, sendo por isso passíveis de serem representados pelo mesmo conceito. Assim, é necessário fazer a sua descoberta para que não existam dois conceitos na mesma ontologia representativos da mesma essência. A terceira etapa desta metodologia é referente à formação de conceitos. Nesta etapa são identificados os conceitos determinados pelos passos anteriores. Segundo Buitelaar, Cimiano e Magnini (2004), para a formação de um conceito deve ser providenciada uma definição intencional de conceito, uma extensão desse mesmo conceito e um conjunto de termos representativos deste. A quarta etapa é relativa à elaboração da hierarquia de conceitos. Nesta etapa, são determinadas as relações taxonómicas entre diferentes classes e subclasses da hierarquia. A quinta etapa diz respeito à descoberta de relações. Nesta etapa é efetuada a determinação e extração das relações não taxonómicas, através de relações encontradas com recurso à utilização de métodos estatísticos e linguísticos, entre outros. Por último, é efetuada a sexta etapa, que consiste na extração de regras ou axiomas. De entre todas as etapas, esta é a menos explorada. Tem como objetivo a elaboração de regras com base no conhecimento obtido através das etapas anteriores, primariamente a partir dos conceitos e relações extraídas (Buitelaar, Cimiano e Magnini 2004) (Al-Aswadi, Chan e Gan 2020).

Após esta descrição sucinta do processo de *Ontology Learning* iremos discutir as diferentes abordagens a este, juntamente com alguns métodos caraterísticos das mesmas. Inicialmente serão explicitados alguns métodos mais antigos pertencentes a abordagens linguísticas ou de *Shallow Learning*, deixando para o final do capítulo a exploração da área de *Deep Learning* e as suas possíveis aplicações em *NLP* e *Ontology Learning*.

3.2 DIFERENTES ABORDAGENS A *Ontology Learning*

As abordagens linguísticas a *Ontology Learning* dependem muitas vezes de ferramentas de *NLP* (processamento de linguagem natural). Estas ferramentas permitem realizar processos de análise de textos a vários níveis, como, por exemplo, a nível semântico e sintático, o que resulta na descoberta de representações de conceitos e relações (Wong, Liu e Bennamoun 2012). Alguns métodos relevantes que resultam desta abordagem são a *Part-Of-Speech (POS) tagging*, a extração baseada em padrões, a análise de estruturas sintáticas e de estruturas de dependências, e o uso de *subcategorization Frames* (SCFs). Estes quatro métodos serão analisados de seguida com mais detalhe.

Comecemos com o método de *POS tagging*. Este utiliza um processo de atribuição de classe e subclasse gramaticais (*POS tag*) a cada palavra de uma frase consoante o contexto em que esta aparece Tran et al. (2009). Este método normalmente auxilia na tarefa de extração de termos e relações (Wong, Liu e Bennamoun 2012). Porém, pode causar algumas ambiguidades (Jurafsky e Martin 2008) (Al-Aswadi, Chan e Gan 2020), especialmente no que toca a palavras homógrafas ou homónimas. Para solucionar este problema, alguns *parsers* são elaborados sobre sistemas de *parsing* estatísticos, o que resulta numa abordagem simultaneamente linguística e de aprendizagem máquina (Al-Aswadi, Chan e Gan 2020) (Wong, Liu e Bennamoun 2012).

A extração baseada em padrões é o processo de reconhecimento de relações através da deteção de padrões em sequências de palavras. Este método é normalmente utilizado para detetar relações de hiperonímia/hiponímia e holonímia/meronímia entre palavras. Porém estas relações não são muito comuns na grande maioria dos textos. Assim, apesar de este método ter uma precisão bastante razoável, os seus valores de *recall* costumam ser bastante baixos (Buitelaar, Cimiano e Magnini 2004) (Al-Aswadi, Chan e Gan 2020) (Wong, Liu e Bennamoun 2012).

A análise de estruturas sintáticas e de estruturas de dependências serve para descobrir novos termos e relações que estejam contidas em frases. No primeiro processo analisam-se estruturas sintáticas para obter termos e relações, enquanto no segundo são utilizadas relações gramaticais para determinar ligações mais complexas (Wong, Liu e Bennamoun 2012). Estes métodos, no entanto, necessitam de cooperar com outros algoritmos para se conseguir uma melhor performance (Al-Aswadi, Chan e Gan 2020).

Uma SCF de uma palavra, normalmente um verbo, é o conjunto de outras palavras numa frase que esta seleciona. Eventualmente a sua análise leva à descoberta das restrições de seleção de cada verbo, que são depois extraídas do texto com o auxílio de *clusters* (Wong, Liu e Bennamoun 2012). Um exemplo de um sistema que aplica esta técnica, juntamente com o uso de *clusters* conceptuais, é o ASIUM (*Acquisition of Semantic Knowledge Using Machine Learning Methods*) (Faure, Nedellec e Rouveirol 1998).

Existem outros sistemas que não utilizam a abordagem referida anteriormente para elaborar ontologias, mas sim uma abordagem mais relacionada com aprendizagem automática. Sistemas que usam este tipo de métodos no âmbito de *Ontology Learning* são denominados sistemas de *Shallow Learning* (Al-Aswadi, Chan e Gan 2020). Nos sistemas de *Shallow Learning* existem dois tipos de métodos possíveis, nomedamente: métodos baseados em estatísticas e métodos baseados em lógica. Vejamos cada um deles.

Segundo Wong, Liu e Bennamoun (2012), os métodos baseados em estatística foram na sua maioria elaborados com a ideia de que a coocorrência de unidades lexicais providencia uma estimativa fiável em relação às suas identidades semânticas. Esta ideia faz com que estes métodos não revelem muita consideração pela semântica e pelas relações entre componentes textuais, o que resulta em que este tipo de métodos seja, normalmente, usado nas fases iniciais de *Ontology Learning*, como a extração de termos ou a construção de relações taxonómicas (Wong, Liu e Bennamoun 2012).

Além dos métodos apresentados existem outros que são dignos de relevo, em particular a análise de coocorrência, o uso de regras de associação, o uso de heurísticas e *clustering* conceptual, a poda de ontologias (*Ontology pruning*) e o uso de *term subsumption*.

Comecemos a revisão destes métodos pela análise de coocorrência. Esta identifica unidades lexicais que tendem a ocorrer juntas para auxiliar na extração de termos relacionados e na descoberta de relações implícitas entre conceitos. Existem medidas de coocorrência que permitem determinar quão forte é a associação entre termos ou entre palavras constituintes de termos multi-palavra (Wong, Liu e Bennamoun 2012). Um exemplo de uma aplicação destas medidas é quando estas foram usadas para efetuar um *clustering* conceptual de palavras após a aplicação de *stemming* (Larkey, Ballesteros e Connell 2002).

Por seu lado, as regras de associação descrevem o nível de ligação entre dois termos, permitindo assim determinar quão forte é a sua relação. A chave para a determinação do nível de abstração vem da aplicação de thresholds como a confiança e o suporte (Wong, Liu e Bennamoun 2012). O suporte indica a fração de *sets* de conjuntos de conceitos (definidas como transações) que incluem os dois conceitos do par a analisar, enquanto a confiança indica a fração de transações que tem um dos conceitos, sabendo que o outro já se encontrava na transação (Maedche e Staab 2000). Estas regras enaltecem as correlações entre palavras-chave nos textos, sendo também fáceis de perceber e interpretar para analistas (Mahgoub 2006).

Quanto às heurísticas e ao *clustering* conceptual, estes são utilizados para agrupar conceitos com base na distância semântica entre eles, que pode ser obtida pela aplicação de diferentes métricas. Um exemplo, já referido nesta dissertação, é o que está descrito em Larkey, Ballesteros e Connell (2002), que usa *clusters* conceptuais elaborados com base em medidas de coocorrência.

A *ontology pruning* permite obter uma ontologia mais pequena, contendo apenas a informação que um utilizador ache relevante a partir de outra ontologia de maiores proporções. Devido ao aumento do conhecimento adquirido em diversas áreas do saber, com o crescimento dos domínios e aumento das quantidades de textos técnicos, o uso de ontologias de grandes dimensões, diante da existência de interesse em apenas uma porção desta, é cada vez mais comum, o que torna este método importante. Alguns exemplos de *ontology pruning* incluem a identificação de conceitos relevantes e a eliminação daqueles que são considerados redundantes (Kim, Caralt e Hilliard 2007).

As métricas de *term subsumption* têm como objetivo indicar quão geral um termo é em relação a outro termo. Quanto mais elevado for o valor desta métrica, mais geral é o primeiro termo em relação ao segundo (Wong, Liu e Bennamoun 2012). Um exemplo do uso deste método é o de Zavitsanos et al. (2007), que identifica conceitos a partir de documentos de texto fornecidos e os organiza numa hierarquia de *subsumption*, descobrindo

tópicos latentes. Também existem casos de uso de uma medida denominada "generalização/especialização", que é baseada em *term subsumption* (Njike-Fotzo e Gallinari 2004).

Os métodos baseados em lógica são menos comuns e normalmente utilizados na obtenção de relações e regras (Wong, Liu e Bennamoun 2012), sendo este tipo de método o mais recomendado no último cenário (Al-Aswadi, Chan e Gan 2020). Existem dois métodos principais que seguem esta abordagem: a programação lógica indutiva (ILP) e a inferência lógica.

A ILP representa um conjunto de métodos que têm como objetivo a derivação de regras positivas e negativas através da coleção de conceitos e relações contida na ontologia. Segundo Boytcheva (janeiro de 2002), os algoritmos de ILP são de especial interesse para aprendizagem máquina, visto que a maioria deles oferece métodos práticos para estender apresentações usadas em algoritmos para solucionar tarefas de aprendizagem supervisionada. Este método, no entanto, necessita de bons *templates* de regras predefinidos por um perito. Um exemplo fornecido por Al-Aswadi, Chan e Gan (2020) é que, se não existirem boas regras negativas, então é possível que seja gerada uma regra inválida.

A inferência lógica tem como propósito a obtenção de relações implícitas entre conceitos a partir de relações já existentes. Estes métodos possuem uma grande possibilidade de gerar relações inválidas ou conflituosas se a intransitividade de algumas relações não for explicitada anteriormente (Wong, Liu e Bennamoun 2012) (Al-Aswadi, Chan e Gan 2020).

Visto que a maior parte das abordagens atrás mencionadas possui caraterísticas próprias, que as tornam mais adequadas a certos passos da construção de uma ontologia, muitas vezes os sistemas de *Ontology Learning* utilizam combinações de métodos destas abordagens. Neste tipo de cenário, que é o mais comum, afirma-se que se utilizou uma abordagem híbrida (Wong, Liu e Bennamoun 2012). Na próxima secção serão discutidos alguns sistemas que utilizam métodos que consubstanciam as abordagens discutidas.

3.3 SISTEMAS DE *Ontology Learning* QUE UTILIZAM AS ABORDAGENS ANTERIORES

Antes de falarmos das aplicações de métodos de *Ontology Learning* é necessário saber como é que o seu sucesso pode ser avaliado. Existem várias maneiras de o fazer. Porém, a grande maioria dos criadores dos sistemas utilizam como medidas de performance a *recall*, a precisão e a *F-measure*. A *recall* indica a fração de objetos relevantes obtidos de forma bem-sucedida dentro do conjunto de termos recolhidos. A precisão indica a fração de objetos recolhidos que são relevantes para o problema proposto (Zhang e Su 2012). A *F-measure* resulta de uma combinação entre a precisão e a *recall*, sendo representada pela fórmula

$$\frac{2PR}{P+R}$$

com P a representar a precisão e R a representar a *recall* (Zhang, Liu e Wang 2016).

O primeiro sistema que será discutido nesta secção é o ASIUM. Este sistema utiliza o *parser* sintático SYLEX para extrair primeiramente *instantiated subcategorization frames*. O método de aprendizagem recebe estas como *input* e divide-se em dois passos: Fatorização, que consiste na elaboração de *basic classes* e

Clustering, que agrega conceitos por *clustering* conceptual. A sua validade depende da qualidade das medidas de similaridade (análise de coocorrência). Ao mesmo tempo *subcategorization frames* agregam *clusters* usando conceitos mais genéricos. Neste programa a intervenção do utilizador ainda é necessária, não só para controlar, mas também para corrigir os *clusters*. Esta é feita através duma interface desenvolvida pelos autores. O utilizador dá nomes aos *clusters* à medida que eles são construídos, o que resulta numa ontologia mais compreensível. O utilizador valida os *clusters* nível a nível. Faure, Nedellec e Rouveirol (1998) afirmaram que *"Intertwining validation and learning in such a way guarantees the relevancy of learned concept while post validation would require Deep revisions to be done by hand"*.

Outro sistema bem conhecido é o CRCTOL (*Concept Relation Concept Tuple Ontology Learning*). Este tem como propósito a extração de conhecimento sob a forma de uma ontologia, a partir de textos de domínio específico. A arquitetura deste sistema semiautomático é composta por três componentes: o componente de *NLP*, um conjunto de algoritmos com funções distintas e um léxico do domínio abordado. O componente de *NLP* usa ferramentas como *POS taggers* e *parsers* sintáticos para processar o texto que é recebido da fase de preprocessamento. Jiang e Tan (2005) afirmaram que isto permite o uso de técnicas de *full text parsing*, ao invés de outras técnicas mais *shallow*. O componente que contém os algoritmos, denominado de *Algorithm Library*, tem um algoritmo estatístico para a extração de conceitos, um algoritmo baseado em regras para extrair relações entre conceitos e um algoritmo de *association rule mining* alterado e generalizado para a construção da ontologia. O léxico com informação respetiva ao domínio abordado é construído manualmente, podendo ser editado ao longo do processo de *Ontology Learning*, para fornecer termos ao componente de *NLP* para este analisar documentos.

O primeiro passo de *Ontology Learning* seguida por este sistema é a relativa ao preprocessamento da informação. Neste passo, o *input* é convertido em texto. De seguida é efetuada a análise *NLP* utilizando *POS tagging* e *tagging* sintática. O passo seguinte é relativo à extração de conceitos, em que os conceitos chave são identificados por um algoritmo estatístico. O penúltimo passo é referente à extração de relações taxonómicas e não taxonómicas entre os conceitos chave do texto. No último passo procede-se à construção da ontologia, interligando os conceitos e as relações encontradas previamente.

Jiang e Tan (2005) efetuaram várias experiências a comparar os sistemas CRCTOL e o Text-To-Onto, um outro sistema de elaboração semiautomática de ontologias. Tais experiências envolveram a extração de termos multi-palavra, a extração de conceitos, a extração de relações semânticas sem verbos auxiliares, e a extração de relações semânticas sem qualquer tipo de restrição. Apesar da *recall* ser muito baixa, quanto à extração de relações semânticas, todas as métricas indicam que a performance do CRCTOL é muito melhor do que a do Text-To-Onto (Jiang e Tan 2005).

O DOODLE e o DOODLE II são sistemas elaborados com o intuito de extrair relações entre conceitos, auxiliando na construção de ontologias. O DOODLE original só ajudava na construção de relações taxonómicas. O DOODLE II ajuda na construção, quer de relações taxonómicas, quer de não taxonómicas, explorando o *WordNet* (base de dados lexical em inglês) e textos de domínio específico, com a auxílio de análise automática de estatísticas de coocorrência lexical e algoritmos de regras de associação. Este sistema é constituído por dois componentes principais: um que extrai relações taxonómicas entre conceitos e outro que extrai relações não

taxonómicas entre os mesmos. O primeiro destes componentes é denominado como *Taxonomic relationship acquisition module* (TRA). Neste o utilizador fornece termos do domínio, algo a que se segue um *spell match* com o *WordNet*, seguido da remoção dos termos internos desnecessários, deixando o modelo *trimmed* que depois é refinado através de duas estratégias: *matched result analysis* e *trimmed result analysis*. O componente que tem como objetivo a extração de relações não taxonómicas é denominado *Non-taxonomic relationship learning module* (NTRL). Este componente extrai os pares de termos que devem estar ligados por alguma relação do texto *corpus*, analisa as coocorrências e aplica o algoritmo de regras de associação — este modelo também extrai candidatos de relações taxonómicas analisando a sua distância no documento. Os modelos do DOODLE II foram testados em textos de domínio específico provenientes de um ramo da área do Direito.

A avaliação do modelo de TRA ficou-se por valores à volta de 30%. Em relação à avaliação do modelo de NTRL, o utilizador avaliou o desempenho dos dois algoritmos utilizados (*Word Space*, que utiliza coocorrência lexical, e regras de associação) em separado e em conjunto. A precisão é muito similar, independentemente do(s) algoritmo(s) utilizado(s), entre 23% e 24%, enquanto a *recall* é também baixo para ambos os algoritmos em separado, subindo, porém, para 57% aquando da sua combinação (Kurematsu et al. 2004).

O sistema HASTI foi elaborado com o propósito de extrair conhecimento. Segundo Shamsfard (2006), as caraterísticas do HASTI são:

- elaborar ontologias sem o conhecimento conceptual e lexical inicial;
- aprender um conjunto muito abrangente de elementos ontológicos, como conceitos, relações taxonómicas e não taxonómicas, e axiomas;
- construir ontologias, de domínio específico ou gerais, dinâmicas, flexíveis a mudanças do utilizador, domínio ou aplicacionais;
- usar uma abordagem híbrida usando vários métodos já referidos anteriormente, como a análise semântica, os métodos heurísticos e as abordagens baseadas em lógica.

Este sistema obtém muito bons resultados com textos simples e restritos e bons resultados em textos reais.

O sistema Hasti é composto por vários módulos. O primeiro é um módulo de *NLP* que utiliza métodos linguísticos para analisar frases de forma morfossintática. Os *outputs* desta fase são novas palavras e *sentence structures* (SSTs). Estas indicam o tema numa frase e serão utilizadas pelo extrator de conhecimento para extrair conhecimento ontológico. O extrator de conhecimento usa métodos de extração de conhecimento a nível textual e frásico, criando elementos ontológicos. Estes são depois colocados na ontologia pelo gestor de ontologias. O ciclo de vida do Hasti tem quatro etapas, podendo estas ser consultadas em detalhe em Shamsfard (2006). A avaliação dos testes foi dividida em duas fases. Numa primeira fase, que consistiu em converter frases em *sentence structures* (SSTs), a precisão e a *recall* foram em média 66% e 76.3%, respetivamente. A precisão aumentou quando a percentagem de palavras novas nas frases diminuiu. Para a segunda fase, na qual se tratou de construir ontologias a partir dos SSTs, os primeiros dois testes foram feitos em textos gerais e os dois últimos foram realizados em textos de domínio específico. No primeiro e terceiro casos a complexidade da gramática nos textos foi restringida, enquanto no segundo e quarto testes foram usados textos reais. Nos primeiros dois

testes foram utilizados textos de domínio específico, enquanto no terceiro e quarto textos foram utilizados textos gerais. A precisão e a *recall* vão diminuindo à medida que se avança desde o primeiro teste até ao quarto. Como seria de esperar, visto que o programa não tem conhecimento conceptual ou lexical no início, o programa obtém muitos melhores resultados em textos mais simples (Shamsfard 2006).

O SERE (*Simultaneous Entities and Relationship Extraction*) é um sistema que, como a sua denominação indica, efetua a extração de entidades e a suas relações de forma simultânea a partir de textos não estruturados. O SERE utiliza *CRFs*, classe de métodos de modelação estatísticos, para a realização destas tarefas. Zhang, Liu e Wang (2016) afirmaram o seguinte: *"Conditional Random Field is a type of discriminative probabilistic model. CRFs are an undirected graphical model (also known as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes"*. O SERE desenvolve o seu trabalho em três passos: 1) a extração de *features* de texto, para determinar a função caraterística; 2) a estimação de parâmetros, que é efetuada ao treinar o modelo com as *features* selecionadas anteriormente para calcular os pesos; e 3) a obtenção e interpretação dos resultados obtidos quando se aplicou o modelo previamente treinado aos dados de teste. As *Deep Belief Networks (DBNs)*, que serão abordadas em maior detalhe neste capítulo, foram usadas em tarefas de *NLP* e produziram resultados satisfatórios. Zhang, Liu e Wang (2016) efetuaram uma comparação entre as *DBNs* e os *CRFs* utilizadas, afirmando que *DBNs* são mais adequadas ao reconhecimento de imagens, ao contrário de *CRFs* que são mais bem aplicadas em tarefas de *NLP*. O SERE foi testado com um *training set* de 80%, e um *test set* de 20%. Zhang, Liu e Wang (2016) concluíram que o SERE pode ser usado para extração de informação de diferentes domínios, afirmando também que este poderia ser melhorado de várias maneiras, como, por exemplo, na capacidade de descobrir relações em contextos mais complexos.

Ao analisar as abordagens anteriormente referidas e os sistemas obtidos através de cada uma delas, identificámos alguns problemas que merecem alguma atenção, nomeadamente o facto de a intervenção humana ser necessária durante todo ou grande parte do processo, a quase inexistência de sistemas para deteção de axiomas, os erros recorrentes de usar *templates* predefinidos para as relações e, principalmente, o facto de os sistemas mencionados anteriormente só trabalharem com pequenos *datasets* ou de domínio específico. Para combater estes problemas muitos engenheiros de conhecimento começaram a tentar elaborar sistemas utilizando *Deep Learning*.

3.4 *Deep Learning* em *Ontology Learning*

Lecun, Bengio e Hinton (2015) definiram uma arquitetura de *Deep Learning* da seguinte forma: *"a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings. Each module in the stack transforms its input to increase both the selectivity and the invariance of the representation. With multiple non-linear layers, say a depth of 5 to 20, a system can implement extremely intricate functions of its inputs that are simultaneously sensitive to minute details . . . and insensitive to large irrelevant variations. . . "*. A atenção dada por Lecun, Bengio e Hinton (2015) ao detalhe torna possível que o *Deep Learning* possa obter melhores resultados que o *Shallow Learning* com o mesmo conjunto

de dados e que possa processar de forma eficiente grandes conjuntos de dados. Em Najafabadi et al. (2015) foi abordada a importância deste tipo de aprendizagem máquina, com o argumento de que o uso de *Deep Learning* implica uma menor interferência do ser humano nos sistemas elaborados, o que é, como já referido, uma das desvantagens do uso de sistemas de *Shallow Learning* para *Ontology Learning*. Os modelos de *Shallow Learning* tradicionais são adequados para a resolução de problemas simples, mas limitados no que toca a algumas aplicações do mundo real, como a linguagem natural. Para estes problemas mais complexos as arquiteturas de *Deep Learning* conseguem produzir resultados melhores (Al-Aswadi, Chan e Gan 2020).

Existem diferentes tipos de redes de *Deep Learning*, sendo estas distinguidas pelo grau de supervisão que lhes é imposto. Estes tipos de redes podem dividir-se em várias categorias: redes supervisionadas, redes não supervisionadas e redes híbridas. Nas redes supervisionadas a informação sobre as *labels* finais está sempre disponível de forma direta ou indireta. Por outro lado, nas redes não supervisionadas não se possui qualquer tipo de informação acerca das *labels* que devem ser obtidas no final. Quanto às redes híbridas, os parâmetros discriminatórios são utilizados para estimar os parâmetros de redes não supervisionadas (Al-Aswadi, Chan e Gan 2020).

Existem muitos tipos diferentes de modelos de *Deep Learning*. Muitos deles possuem propósitos específicos, porém podem ser alterados, ou até combinados, com outros tipos de modelos de *Deep Learning* para servirem outros propósitos. De entre os vários tipos de modelos existentes abordaremos apenas quatro nesta secção: Redes Neuronais Recorrentes (RNNs), Redes Neuronais Convolucionais (CNNs), *Deep Belief Networks* (DBNs) e *Autoencoders*.

As RNNs tanto podem ser supervisionadas como não supervisionadas. Usualmente, são utilizadas para prever informação futura com base em sequências passadas de informação. Por conseguinte, em tarefas que envolvam *inputs* sequenciais, as RNNs costumam ter uma boa performance (Al-Aswadi, Chan e Gan 2020). Segundo Lecun, Bengio e Hinton (2015), as RNNs são sistemas muito poderosos. Contudo, o seu treino tem provado ser bastante problemático, visto que os gradientes de *backpropagation* ou crescem ou diminuem em cada passo. Logo, ao longo de muitos passos, esses gradientes tendem a aumentar ou a diminuir muito até atingirem quantidades ínfimas. As RNNs convencionais possuem nodos de *input*, nodos escondidos e um nodo de *output*. Cada nodo escondido recebe informação de um nodo de *input*, aplica-lhe um algoritmo e envia-a para um outro nodo escondido. Neste tipo de redes existem três tipos de parâmetros (Le 2015): 1) o representativo do valor transmitido entre uma camada de *input* e uma camada escondida; 2) o representativo do valor transmitido entre duas camadas escondidas; e 3) o representativo do valor transmitido entre uma camada escondida e a camada de *output*. Apesar de o principal propósito das RNNs ser a aprendizagem de dependências a longo prazo, é difícil aprender a armazenar informação durante uma grande quantidade de tempo. Para resolver este problema foi sugerido fornecer uma memória explícita à rede. Com isso, em primeiro lugar, surgiu a rede LSTM, que usa unidades escondidas e especializadas, denominadas células de memória, cujo propósito é memorizar informação durante uma grande quantidade de tempo. As redes LSTM provaram desde então serem mais eficientes que as RNNs convencionais (Lecun, Bengio e Hinton 2015). Os modelos baseados em RNNs já foram utilizados para fazer a extração de ontologias a partir de texto, por exemplo, em Petrucci, Ghidini e Rospocher (2016), que produziram um sistema que transformava frases de um tipo específico (definições) em axiomas OWL.

As CNNs são redes multi-camada supervisionadas e *feed-forward*. Estas redes têm tido muitos tipos de aplicações, principalmente em reconhecimento de imagem e vídeo, mas também em *NLP*. Estas redes possuem as desvantagens de terem um custo computacional muito elevado e necessitarem de muitos *datasets* de treino (Al-Aswadi, Chan e Gan 2020). Segundo Lecun, Bengio e Hinton (2015), a arquitetura das CNNs está dividida em várias fases, com as primeiras fases a serem compostas por camadas convolucionais e de *pooling*. As camadas convolucionais têm como propósito detetar conjuntos de *features* provenientes da camada anterior e estão organizadas por mapas de *features*, em que cada unidade se encontra ligada à camada anterior através de um conjunto de pesos, denominado *filter bank*. As camadas de *pooling* têm como propósito fundir termos semanticamente similares. Uma unidade de *pooling* analisa o máximo de unidades de um ou vários mapas de *features*. As outras unidades de *pooling* só recebem *inputs* de *patches* a mais de uma linha e coluna de distância, reduzindo a dimensão da representação e criando uma resistência a pequenas variações e distorções (Lecun, Bengio e Hinton 2015). A seguir a várias fases que envolvem combinações de diferentes instâncias das camadas anteriormente referidas, chegamos a uma ou duas camadas escondidas e totalmente conectadas que conectam o *output* da última camada anterior ao classificador (Le 2015). Estas camadas incluem parâmetros que necessitam de computação complexa nos exemplos de treino, o que leva a que seja preciso eliminar nodos e conexões, o que normalmente se faz usando uma técnica de *Dropout* (Albawi, Mohammed e Al-Zawi 2018). Como já foi referido anteriormente, as CNN são redes maioritariamente aplicáveis a vídeo e a imagem. Porém este tipo de redes é ocasionalmente utilizado noutras áreas, nomeadamente em *NLP*. Um problema que as CNNs têm, relacionado com esta área, é que estas precisam de várias camadas convolucionais para capturar dependências a longo prazo, devido à pequena dimensão local destas e das camadas de *pooling* (Hassan e Mahmood 2018). Este facto não impede, no entanto, que alguns modelos entretanto desenvolvidos utilizem CNNs. Kim (2014), por exemplo, utilizou um modelo que possui apenas uma camada convolucional, em cima de vetores de palavras, pré treinados com a arquitetura de CBOW *(Continuous Bag-Of-Words)*, atingindo resultados positivos. Este afirma que os resultados obtidos comprovam que o pré treino não supervisionado de vetores de palavras é um importante ingrediente para o *Deep Learning* envolvendo *NLP*. Outro exemplo da aplicação de CNNs nesta área é o de Shen et al. (2014), onde são utilizadas uma camada convolucional e uma de *pooling* para fornecer *features* representativas do contexto a outras camadas que fazem *matching* semântico entre documentos e *queries*.

Em relação a processamento de linguagem natural as CNNs possuem um problema claro relacionado com o facto de o seu *input* não poder ser variável, ao invés das RNNs. Pode-se tentar lidar com este problema ao aplicar uma camada de *pooling* a todo o *output*, fazendo com que, o *output* tenha um tamanho fixo, apesar de o *input* ter um tamanho variável. O problema com esta abordagem é que o *output* dos neurónios de *max pooling* é invariante à translação, o que é desejado para diminuir a distorção em imagens, porém nesta caso resulta em que num grande *max pooling* a perda de informação relativa à posição seja inevitável, sendo esta uma desvantagem do uso de CNNs face a RNNs (Le 2015).

Como pudemos observar com os exemplos mencionados anteriormente de RNNs e CNNs, é comum a modificação de modelos para a sua otimização, consoante as circunstâncias em que o modelo vai ser utilizado. Existem inclusive modelos que foram criados com a junção destes tipos de redes, como é o caso de Hassan e Mahmood (2018), que efetuam extração de termos estabelecendo depois relações entre estes. A arquitetura

deste modelo envolve quatro tipos de camadas: uma camada de *input* que usa *word embeddings*, uma camada convolucional, uma camada recorrente LSTM e uma camada de classificação. Hassan e Mahmood (2018) não utilizaram nenhuma camada de *pooling*, pois estes acreditam que este tipo de camada é responsável pela perda de informação a nível local, visto que só captura as caraterísticas principais de uma frase, ignorando as outras.

As DBNs (*Deep Belief Network*) são redes *feedforward*, que tanto podem ser supervisionadas, como não supervisionadas. Estas são constituídas por uma rede de *Restricted Boltzmann Machines* (RBM). Tal como as CNNs, as *DBNs* têm sido utilizadas em processos de reconhecimento de vídeo e imagem, tendo também sido aplicadas em *NLP*. Al-Aswadi, Chan e Gan (2020) afirmaram que, quando um *dataset* é da área da Visão por Computador, as CNNs costumam ter melhor performance que as *DBNs*. Porém, quando o *dataset* não é dessa área, as *DBNs* podem definitivamente obter resultados melhores. Em relação à estrutura deste tipo de redes, cada par de camadas forma uma *RBM (Restricted Boltzman Machine)*, ou seja, cada par de camadas está totalmente interligado sem que existam quaisquer ligações entre elementos da mesma camada. Cada camada visível está ligada à camada escondida da *RBM* anterior e as duas camadas de cima são não direcionais. A direção das conexões entre as camadas é feita *top-down*. O treino das *RBMs* constituintes é sequencial, com as *RBMs* mais em baixo a serem treinadas primeiro que as mais em cima. As *features* são extraídas pela *RBM* mais acima e são, de seguida, propagadas para as camadas mais abaixo (Liu et al. 2017). O treino deste tipo de redes pode ser dividido em duas fases: a de pré-treino e a de *fine-tuning*. Na primeira fase, um treino não supervisionado é efetuado *bottom-up* para extrair *features*. Na segunda fase, um treino supervisionado é efetuado *top-down*, para ajustar os valores do modelo. As *DBNs* foram utilizadas já em várias áreas relacionadas com o processamento de linguagem natural, como, por exemplo, no modelo criado em Zhong et al. (2016), denominado EAEDB (*Entity Attribute Extraction Based on Deep Belief Network*), que inclui uma *DBN* com o intuito de extrair informação, mais concretamente, entidades e atributos de textos não estruturados em mandarim.

Uma rede *Autoencoder* é uma rede *feed forward*, que tanto pode ser supervisionada como não supervisionada (Al-Aswadi, Chan e Gan 2020). Segundo Alom et al. (2019), o principal objetivo de uma rede *Autoencoder* é aprender e representar codificações de um *input*, tipicamente para redução, compressão ou fusão das dimensões da informação, entre outras operações. A técnica do *Autoencoder* divide-se em duas fases. A primeira fase é relativa ao *encoding*. Segundo Alom et al. (2019), as amostras de *input* são mapeadas num *feature space* mais pequeno, sendo este processo repetido até que este fique com as dimensões desejadas. A segunda fase é relativa ao *decoding*, no qual as *features* de pequenas dimensões são regeneradas seguindo o processo inverso. O modelo é treinado no sentido de minimizar os erros de reconstrução, que são definidos por uma função de *loss*. Quando se trata de *autoencoding* multi-camada, a mesma operação é repetida as vezes que forem necessárias, sendo os processos de codificação e descodificação estendidos ao longo de várias camadas escondidas.

Existem alguns tipos diferentes de modelos baseados em *Autoencoders*, como, por exemplo, o *Variational Autoencoder* (VAE), que se encontra explicitado em Xie e Ma (2019) e em Shao et al. (2020). *Variational Autoencoders* costumam ser aplicadas a imagens, porém, estas também podem ser aplicadas a área relacionadas com o processamento de linguagens, como aconteceu em Xie e Ma (2019) e Shao et al. (2020). Xie e Ma (2019) propuseram um novo sistema, que denominaram de *Dual-View Variational Autoencoder*. Este é um sistema que foi utilizado para estabelecer uma correspondência entre pares de frases e que tem o nome de *Dual-View*, (ou

Dupla-Visão), porque unifica duas abordagens diferentes para o fazer: uma abordagem através de *embedding* e uma abordagem por interação. Por sua vez, Shao et al. (2020) apresentaram um sistema denominado *ControlVAE* para melhorar a performance de *VAEs* já existentes, ao estabilizar o valor da divergência entre a distribuição aproximada aprendida e o valor real das variáveis.

Parte II

COMPONENTE PRÁTICA

4

O CASO DE APLICAÇÃO

4.1 APRESENTAÇÃO GERAL

O principal objetivo desta dissertação envolveu a extração e análise da informação contida nos diversos fólios constituintes do Tombo da Mitra, de forma que seja possível a elaboração de uma ontologia que torne esta informação mais compreensível e organizada. Em si este objetivo não aparenta apresentar grandes dificuldades, porém existem alguns obstáculos que tornam esta tarefa mais difícil de ser concluída.

É necessário relembrar que os textos contidos nos fólios do Livro das Propriedades foram concebidos nos inícios do século XVII, pelo que o português presente nestes possui diferenças bastante acentuadas em relação ao português atual, existindo mesmo a natural variação ortográfica, mas também fonética, morfossintática, semântica e lexical, no português clássico utilizado nos diferentes fólios. Este facto causou algumas dificuldades na utilização de algumas metodologias testadas, maioritariamente relacionadas com diferenças na grafia ou no significado dos termos. Também é preciso notar que é necessário analisar centenas de fólios de tamanho grande, o que torna essencial que o processo seja automatizado.

Obter estas informações acerca das propriedades fez com que fosse possível observar como os terrenos se encontravam divididos pelo arcebispado, a quem se encontravam emprazados ou que fins serviam, possuindo, por conseguinte, grande relevância histórica. Para estruturar esta informação, foi elaborada uma ontologia, que, aquando da sua consulta, permite perceber as interações existentes entre os diferentes elementos constituintes do domínio abordado pelos fólios. As diversas instâncias de propriedades são analisadas e os seus aspetos são categorizados, de forma a representar a estrutura descrita nos fólios da forma mais fidedigna possível. Porém esta ontologia não permite só a estruturação das propriedades. Outros aspetos relevantes que permite consultar envolvem os títulos a que pertenciam as propriedades, os emprazadores ativos em cada título e até os pagamentos forais efetuados.

Por último, é importante referir que a ontologia, não só permitiu estruturar a informação existente nos fólios, como também permitiu inferir outros aspetos, passíveis de se depreenderem a partir da informação categorizada. Após a elaboração da ontologia, foi possível calcular, por exemplo, a área da maioria dos terrenos, a partir das dimensões recolhidas, o que tornou possível obter a área mínima dos terrenos de títulos e até do arcebispado.

4.2 MÉTODOS TESTADOS E UTILIZADOS

O conteúdo dos fólios de propriedades que foi analisado é de domínio específico, possuindo, por conseguinte, um vocabulário muito caraterístico da descrição de diferentes aspetos de terrenos e propriedades, tais como as suas dimensões ou o seu valor em termos de cultivo, entre outros. Apesar do conteúdo dos fólios não ser estruturado, este não está completamente desprovido de padrões. Existem várias expressões que aparecem repetidamente ao longo dos vários fólios, quase sempre com o mesmo propósito, introduzindo diferentes aspetos dos terrenos registados ou mesmo os próprios terrenos. Devido à complexidade da implementação de uma metodologia de *Deep Learning*, e como os textos são de domínio específico, possuindo padrões que se repetem regularmente ao longo do texto, foram utilizadas metodologias de processamento de linguagens e aprendizagem máquina, em vez de redes de *Deep Learning*. O processo de elaboração da ontologia para o Livro das Propriedades seguirá os passos requeridos pela metodologia *Ontology Learning Layer Cake*, com algumas adaptações (Buitelaar, Cimiano e Magnini 2004). De seguida, apresentaremos o modo como cada um dos seus passos foi executado.

4.2.1 *Extração de termos*

O primeiro passo da metodologia envolve a extração de termos considerados relevantes para a ontologia. Este passo é de grande importância, pois tem um peso bastante elevado nos termos constituintes da ontologia, não só relativamente a conceitos, mas também no que concerne a relações e atributos. No nosso processo foram utilizados alguns métodos diferentes, que nos propiciaram resultados distintos, a saber:

Termos mais comuns - Descrição do método

O primeiro método testado fez a extração dos termos mais comuns num texto. Este assumiu que os termos que ocorrem mais vezes num texto são os mais relevantes. Este é um método bastante simples. Primeiramente, foi necessário obter todas as palavras integradas nos fólios à nossa disposição, tendo sido utilizadas para este efeito as bibliotecas *os* (*OS Library*) e *os.path* (*OS.Path Library*). O conteúdo dos fólios ficou armazenado num *array*. De seguida, visto que este método não tem em consideração o contexto dos termos, foram removidos todos os sinais de pontuação do texto, uma vez que alguns destes, como vírgulas ou pontos finais, influenciariam os resultados finais, devido a serem elementos bastante comuns. O passo seguinte envolveu a remoção de *stop words*. Uma *stop word* é uma palavra utilizada de forma comum, que não acrescenta nenhum valor a um texto em termos de contexto, sendo independente deste e que, por conseguinte, tem que ser ignorada. Este tipo de palavras inclui determinantes ou pronomes pessoais ou algumas preposições, entre outros. O *set* de *stop words* que foi utilizado pertence à biblioteca *NLTK* (*NLTK Library*). Visto que este conjunto contém todas as palavras sem maiúsculas, as palavras contidas nos fólios foram passadas para *lowercase* para que fosse possível retirar as *stop words*. Após a sua remoção, foi criado um dicionário, no qual a palavra foi o índice e a contagem do número de vezes em que esta aparecia nos fólios o seu valor. Posteriormente, este dicionário foi ordenado de forma decrescente consoante o valor do *array*.

Termos mais comuns - Resultados obtidos e problemas encontrados

Os principais termos que foram obtidos através deste método podem ser consultados no Anexo A desta dissertação.

Ao observar estes resultados podemos identificar alguns problemas. O principal problema que encontrámos foi a presença de palavras que habitualmente estão contidas dentro do grupo de *stop words*. Porém, devido ao facto de o português abordado ser do século XVII, estas estavam escritas de maneira diferente, não sendo por isso detetadas pelo método. Exemplos destes tipos de palavras são "hua" e "co", que correspondem, respetivamente, à palavra "uma" e à palavra “com” no português atual.

Existem também outros termos que aparecem repetidos, por causa de aparecerem no singular e plural, como são os casos das palavras "terra" e "terras".

O primeiro problema mencionado foi a maior dificuldade que encontrámos no processo de elaboração da ontologia pretendida, não sendo possível utilizar textos em português atual, mas apenas em português clássico. Isto foi feito com o intuito de a ontologia já se encontrar preparada para lidar com português clássico. Desta forma, podemos concluir que é extremamente difícil obter somente termos relevantes.

Uso da ferramenta YAKE - Descrição do método

Outro dos métodos que foram testados baseou-se na utilização da ferramenta YAKE (Campos et al. 2018) (Campos et al. 2020) para fazer a extração de termos relevantes. Esta ferramenta foi encarregue de extrair as palavras que considerasse mais relevantes consoante os parâmetros que lhe foram fornecidos. O primeiro parâmetro deste extrator a ser definido foi a língua, que, obviamente, foi definida como português. De seguida, foram efetuadas duas experiências: uma, com o intuito de obter termos compostos só por uma palavra, e uma segunda, para obter termos com duas palavras, para verificar se algum dos termos constituídos por duas palavras é relevante para a ontologia. Por fim, foi definido o número de termos extraídos, 20. Como o extrator foi aplicado de forma isolada a cada um dos fólios, foram extraídos no total 20 termos por fólio.

Uso da ferramenta YAKE - Análise dos resultados obtidos

No anexo B encontram-se duas tabelas contendo os resultados da aplicação do YAKE a dois dos fólios do Tombo da Mitra.

Ao consultar esses resultados, podemos verificar que são mais completos dos que os resultados anteriores, em particular em termos de instâncias. No entanto, ainda podemos encontrar alguns termos que não acrescentaram informação à ontologia, por exemplo, as palavras "uai pera", que no português atual correspondem a "vai para". Porém, este tipo de termos não foi extraído de forma comum, sendo a maior parte deles relevante. O YAKE também conseguiu discernir, na grande maioria dos casos, as *stop words* presentes nos fólios, não as incluindo mesmo quando estas se encontravam em português clássico. Apesar desta metodologia ter conseguido resultados superiores, não é sensato descartar todos os resultados obtidos com o algoritmo de termos mais comuns, visto que este conseguiu extrair alguns termos relevantes, como foi o caso "parte" ou "semeadura", entre outros, que serão abordados aquando da consulta do Vocabulário de Bluteau (Barros 2018) (Bluteau 1712-28). O

relativo sucesso do primeiro método não se deveu somente à remoção das *stop words*, mas também ao facto de o conteúdo dos fólios ser de um domínio específico, domínio esse que, neste cenário, se encontra associado à descrição das propriedades referidas no Tombo da Mitra. Textos de um domínio específico tendem a ter um vocabulário mais reduzido, o que leva a que os mesmos termos apareçam em várias ocasiões ao longo dos fólios, acentuando a contagem do número de vezes em que estes ocorrem. Esta é a razão pela qual estes termos se destacaram de forma considerável no número de vezes em que são utilizados nos fólios.

No YAKE, foi obtida uma maior variedade nos termos extraídos do que nos termos resultantes da procura dos termos mais comuns, em particular no que toca a instâncias (na sua grande maioria nomes de pessoas), como se pôde verificar nos resultados da aplicação do YAKE ao fólio 241. A possibilidade de extrair termos multipalavra também facilitou a captação de instâncias, visto que a grande maioria das instâncias consideradas relevantes pelo YAKE são termos multipalavra, sendo a maior parte dos termos de uma só palavra conceitos, relações ou até mesmo atributos. Isto foi constatado durante a realização dos próximos passos do processo de construção da nossa ontologia.

Devido aos resultados da extração dos dois métodos serem relevantes, a próxima etapa foi aplicada aos *outputs* de ambos os métodos, com o intuito de obter os melhores resultados possível e de investigar se os mesmos poderiam ser complementados.

Tendo sido efetuada a extração de termos relevantes para a ontologia, foi possível prosseguir com o passo seguinte da metodologia, que, segundo o *Ontology Learning Layer Cake*, seria a deteção de sinónimos. De facto, no que toca ao *output* dos termos obtidos pela contagem dos termos mais comuns, o próximo passo a realizar seria a deteção de sinónimos. Porém, como os termos obtidos através do YAKE possuíam uma percentagem significativa de instâncias, e as instâncias não possuíam sinónimos, foi efetuada uma separação dos termos, para que a deteção dos sinónimos não fosse aplicada às instâncias. Tal circunstância fez com que a etapa de deteção de sinónimos não fosse uma etapa dependente da etapa relativa à formação de conceitos nesta situação.

Apesar deste último ponto, a deteção de sinónimos será discutida de seguida. Mais à frente na ontologia será referida quando esta foi aplicada aos termos extraídos pelo YAKE.

4.2.2 *Deteção de sinónimos*

Numa ontologia não podem existir elementos com a mesma essência ou que representem os mesmos conceitos. É normal que, ao longo dos fólios, sejam utilizados diferentes termos para referir um conceito ou relação. Isso faz com que seja necessário escolher um destes termos para representar o conceito, de forma que a ontologia fique corretamente elaborada.

Descrição do método

Para a realização deste método foi consultada uma *API (Application Programming Interface)* pública (*Dicionário Online de Português*). Para que pudéssemos utilizar esta *API* foi necessário efetuar um pedido *GET*, com o *link* da *API* e a palavra da qual se quer saber a lista de sinónimos, o que tornou necessária a utilização da biblioteca *Requests* (*Requests Library*). Após a realização do referido pedido *GET*, a informação recebida

continha muito mais do que a lista dos sinónimos da palavra, devido a esta estar em HTML. Assim, foi necessário filtrar todos os dados que não fazem parte da lista de sinónimos, uma vez que são considerados irrelevantes neste contexto. Para efetuar este filtro foi utilizada a biblioteca *Beautiful Soup* (Richardson 2016), que foi desenvolvida com o propósito de retirar informação a partir de ficheiros HTML e XML. Após a obtenção da pequena secção de HTML que continha a lista de sinónimos da palavra pedida, foi necessário ainda remover as *tags* que estavam nesta lista. Para cumprir tal objetivo foram utilizadas expressões regulares.

Resultados obtidos

Para testar o bom funcionamento da *API*, aplicámo-la à parte ao termo "terra", com os seguintes resultados:

```
sinonimos_terra = get_sinonimos("terra")
print(sinonimos_terra)

Output: ['terreno']
```

[terra]. In: LEXICO, Dicionário Online de Português. Porto: 7Graus, 2018.
Disponível em: [http://www.lexico.pt/terra]. Acesso em: 15/07/2021.

Análise e validação dos resultados

A *API* que utilizámos encontra-se funcional, sendo capaz de apresentar resultados relevantes para a ontologia a desenvolver. Ao observar o exemplo anterior, constatamos que o termo "terreno" foi detetado como sinónimo de "terra", o que é uma assunção válida no contexto dos fólios, pois os "terrenos" da Mesa Arcebispal são as suas "terras". Assim, agora no desenvolvimento da ontologia será utilizado este último termo sempre que se fizer menção a "terreno", visto que é o termo mais comum de entre os dois.

Embora a operacionalidade da *API* seja boa e os seus contributos (*outputs*) sejam relevantes, identificámos alguns aspetos na sua utilização não muito positivos. O principal aspeto foi, novamente, o facto de não ser possível encontrar sinónimos para uma parte dos termos contidos nos fólios, visto que a sua ortografia se encontra diferente no português atual, em comparação com o português clássico. Este problema não pôde ser ultrapassado, devido a não ter sido encontrado um dicionário ou *API* que tivesse em consideração o português do século XVII e o alto grau de complexidade que se encontra envolvido na elaboração de um dicionário de sinónimos próprio para todos os termos presentes nos fólios.

Apesar destes problemas, podemos concluir que o uso desta *API* tem uma influência bastante positiva na ontologia ao possibilitar a execução desta etapa de forma competente.

De seguida, será discutida a etapa de formação de conceitos, quer para os termos mais comuns, quer para os termos provenientes da ferramenta YAKE.

4.2.3 *Formação de Conceitos*

Nesta etapa inicia-se a separação dos termos, que foram obtidos pela extração e reduzidos pela deteção de sinónimos, em elementos distintos da ontologia, tendo em especial atenção os conceitos. Primeiro será abordada a aplicação desta etapa aos termos obtidos pela contagem dos termos mais comuns, sendo depois feita a aplicação do trabalho realizado nesta etapa aos termos obtidos a partir da extração efetuada pelo YAKE na primeira etapa.

Aplicação aos termos obtidos através da contagem — Descrição do método

Esta etapa tem como função obter os conceitos mais relevantes da ontologia. Este tipo de elementos tende a possuir a sua própria essência, simbolizando algo de forma concreta ou abstrata o que, segundo Hejl (2014), é representado na sua maioria por nomes ou substantivos comuns. De forma a usufruir desta caraterística foi efetuada *Part Of Speech Tagging (POS Tagging)*, a fim de determinar os nomes comuns de entre todos os termos restantes das etapas anteriores. Para efetuar o *POS Tagging*, foi utilizado o *Tagger* do (*Polyglot*), devido à sua aptidão com línguas que não a inglesa. Após a aplicação do (*Polyglot*) aos termos e da obtenção de termos com a *POS Tag* de "*NOUN*" que representa um nome comum, foi criado um dicionário com os 100 termos mais comuns ordenados, do mais frequente para o menos frequente. Estes termos foram depois armazenados num ficheiro CSV.

Aplicação aos termos obtidos através da contagem — Resultados obtidos (os 20 conceitos mais comuns, obtidos por este método)

A lista dos resultados obtidos é composta pelos termos: "parte", "terra", "semeadura", "norte", "varas", "hua", "alqueires", "caminho", "terras", "casal", "he", "partes", "meo", "campo", "casas", "propriedades", "lugar", "casa", "en" e "prazo".

Nestes resultados ainda foi possível constatar alguns termos provenientes da primeira etapa, que devido às diferenças de ortografia entre o tipo de português dos fólios e o português atual não foram consideradas *stop words*. De entre estes 20 termos os que se encaixam nessa categoria são "hua", "meo", "he" e "en".

Removendo estes 4 termos e acrescentando os 4 termos que não representam *stop words* seguintes a lista fica assim: "parte", "terra", "semeadura", "norte", "varas", "alqueires", "caminho", "terras", "casal", "partes", "campo", "casas", "propriedades", "lugar", "casa", "prazo", "parede", "camara", "monte" e "dia".

Aplicação aos termos obtidos através do YAKE — Descrição do método

Como foi referido anteriormente, foi efetuada uma separação dos termos extraídos pelo YAKE, para que a aplicação da deteção de sinónimos não fosse aplicada a instâncias. Para efetuar esta separação foi utilizado, tal como no método anterior, o (*Polyglot*) para efetuar o *POS Tagging* dos termos extraídos. O algoritmo elaborado teve em atenção a possibilidade de os termos serem multipalavra, visto que muitos dos termos extraídos na primeira etapa pelo YAKE eram binómios. Instâncias como nomes de pessoas ou de organizações costumam ser

representadas por nomes próprios. Assim, se a *POS Tag* correspondesse à de um nome próprio ("PROPN"), o algoritmo considerava-a uma instância, ou uma palavra de uma possível instância quando se tratasse de um termo multipalavra. Quando o termo é multipalavra, para ser considerado um conceito, todas as palavras constituintes têm que ter como classe gramatical nome comum, enquanto para o termo ser considerado uma instância, este tem que ser constituído exclusivamente por nomes próprios. Se houver uma palavra no excerto não correspondente a um nome comum ou a um nome próprio avança-se para a expressão seguinte, sendo que excertos que possuem uma combinação de nomes próprios e nomes comuns são também ignorados, havendo duas razões para ser tomada essa decisão. Primeiramente, a maior parte dos nomes que aparecem em binómios aparecem noutros excertos, sendo provável que a eliminação não resulte em perda de informação. Por exemplo, antes de ser efetuada esta decisão, quer "Mesmo Afonso", quer "Afonso Esteues" apareciam como possíveis elementos ontológicos, sendo que com este filtro "Mesmo Afonso" deixa de ser considerado, visto que "Mesmo" não é um nome próprio. O segundo ponto é que binómios que contêm nomes próprios e comuns não podem ser classificados de forma segura como conceitos ou instâncias, visto que não se pode dar prioridade a nenhuma das classes. Uma conclusão a que foi possível chegar foi que uma elevada percentagem dos binómios que são considerados instâncias são nomes de pessoas, na sua maioria emprazadores das propriedades da Mesa Arcebispal. Por conseguinte, estes *outputs* foram armazenados num ficheiro CSV separado dos outros termos considerados instâncias, para uma possível futura utilização.

Após a separação dos termos extraídos pelo YAKE, em conceitos e instâncias, a deteção de sinónimos foi aplicada à lista de conceitos obtida anteriormente. A lista de conceitos e a lista de sinónimos são percorridas e caso o sinónimo exista no dicionário este é eliminado. A lista de conceitos final é depois armazenada num ficheiro CSV.

Aplicação aos termos obtidos através do YAKE — Resultados obtidos

A lista de conceitos obtida pelo YAKE foi: "norte", "sul", "poente", "semeadura", "largo", "campo", "caminho", "terra", "varas", "casal", "terras", "herdade", "herdades", "dito", "elle", "tombo", "juramento", "dita", "hum", "propriedades", "casas", "min", "moinho", "rega", "alqueires", "casa", "estrada", "meo", "villa", "parede" e "monte".

Apesar de ainda existirem termos presentes nesta lista que não são conceitos, sendo que alguns deles só se encontram presentes devido às diferenças de ortografia já mencionadas (como "hum"), os resultados são positivos, só existindo os seguintes termos que não podem ser considerados conceitos: "dito", "elle", "dita", "hum", "min" e "meo", (quando é representativo de "meu"); sendo, por conseguinte, estes termos removidos da lista de conceitos.

Uma comparação entre resultados obtidos pelos dois métodos referidos anteriormente encontra-se apresentada no Anexo C.

Análise e validação dos conceitos, relações e atributos

Para compreender melhor quais os conceitos suficientemente relevantes para serem elementos da ontologia é necessário verificar quais as suas definições. Isto é feito com o intuito não só de identificar a maneira como estes se ligariam com outros conceitos, mas também se estes pertencem de facto a esta ontologia na

qualidade de conceitos ou como outra espécie de elementos ontológicos. Porém, para obter as definições dos elementos desta lista não se podem consultar dicionários atuais da língua portuguesa, não só devido à evolução semântica das palavras, como também devido a alterações nas definições das próprias palavras. Para contornar estes obstáculos consultámos o Dicionário de Bluteau, elaborado por Rafael Bluteau e publicado entre 1712 e 1728 (Bluteau 1712-28), sendo as definições existentes nos volumes deste dicionário uma representação mais fidedigna do significado dos elementos constituintes da lista. Estes, acompanhados das respetivas definições, encontram-se apresentados no anexo D.

Após consultar as definições presentes no Vocabulário de Bluteau para estas palavras, foi possível obter algumas conclusões acerca dos elementos da lista obtida nesta etapa, que se apresentam de seguida.

- "norte", "sul" e "poente" — são todas orientações que, juntamente com "nacente" ou "nascente" são utilizadas para indicar confrontações entre diferentes propriedades, sendo por isso representativas dessas confrontações (também podendo ser denominadas de fronteiras), conectando duas propriedades. Esta conclusão leva a que estes termos sejam mais adequados como representantes de relações, ligando diversas propriedades com base nas confrontações existentes entre elas, tendo em consideração a(s) orientação(ões) em que se encontram essas confrontações.

- “semeadura” — diz respeito ao cultivo das terras. Nos fólios este termo é utilizado regularmente na expressão "levara de semeadura", com o intuito de indicar a quantidade de semente, ou seja, de cereal, necessária para semear (em) toda essa terra. Logo, a "semeadura"ou "semente", como também se diz no Livro, traduz o valor da terra, o seu tamanho e produtividade. Consequentemente, é mais lógico que este termo não seja considerado um conceito com essência própria, mas sim uma caraterística de uma terra, o que faz com que este seja um atributo dessa mesma terra.

- “largo” — no português atual equivale à palavra "largura". O sentido em que este termo é utilizado para indicar a largura de um terreno. O comprimento de um terreno é mencionado de forma tão regular como a largura ao longo dos fólios, sendo escrito como "comprido" nestes. É por isso possível que o POS Tagger do Polyglot tenha considerado este termo como um adjetivo, o que explica a razão de este não aparecer nesta lista. Estes dois termos são representativos das dimensões de um terreno, sendo esta uma caraterística desse mesmo terreno. Por esta razão, é passível de ser concluído que a largura e comprimento do terreno constituem um atributo composto da propriedade, que pode ser denominado de "dimensões".

- “caminho” - representativo de caminhos que ligam propriedades distintas, aparecendo por isso regularmente aquando da apresentação das confrontações.

- “campo” - representativo de um tipo de terra, podendo ser simplesmente uma instância de um possível atributo tipo.

- “terra” - representativo de um terreno pertencente à Mesa Arcebispal. Este é um termo com essência própria, não sendo uma caraterística ou estabelecendo uma relação entre outros termos, pelo que se pode considerar este termo como um conceito.

- “varas” - medida utilizada de forma mais regular ao longo dos fólios para medir o comprimento e a distância das terras.
- “casal” - utilizado para descrever uma propriedade rural com casas, de uma ou mais famílias.
- “terras” - plural do termo terra, pelo que não é necessário discuti-lo.
- “herdade” - tipo de propriedade ou terra podendo eventualmente ser uma instância de um atributo "tipo" desta.
- “herdades” - plural do termo *herdade*, pelo que não é necessário explicitá-lo.
- “tombo” - relativo ao Tombo da Mitra em si, ou a outros livros onde eram tombadas ou registadas propriedades.
- “juramento” - não é relevante para a ontologia, visto que diz respeito aos juramentos dos escrivães nos fólios iniciais.
- “propriedades” - diz respeito aos terrenos da Mesa Arcebispal, podendo ser interpretado como um sinónimo de terras neste cenário. Este sinónimo não foi considerado como tal na etapa de deteção de sinónimos, visto que em circunstâncias normais estes dois termos não o são.
- “casas” e “moinho” - dizem respeito a outros possíveis tipos de propriedades.
- “rega” - normalmente é utilizado aquando da avaliação de um terreno para determinar se este tem água (para efetuar a rega), com o intuito de determinar se este é próprio para certos tipos de cultivo.
- “alqueires” - representativo de uma pequena caixa que era utilizada para armazenar e medir a quantidade de cereal disponível. Esta era a medida para determinar quanto cereal era possível semear em cada terra.
- “casa” - singular de *casas* tendo o mesmo significado que este termo.
- “estrada” - sinónimo de *caminho* nestes fólios. Tal como foi referido anteriormente em relação aos termos *terras* e *propriedades*, esta não foi detetada como sinónimo de *caminho*, devido ao facto de a sinonímia destas palavras depender do contexto em que são utilizadas.
- “villa” - indicativo de um tipo de localidade. Este termo não é necessário para a ontologia, visto que localidades raramente são mencionadas nos fólios, não havendo razão para estas constituírem um elemento da ontologia.
- “parede” - representativo dos muros que cercavam alguns dos terrenos mencionados nos fólios, não sendo um elemento relevante da ontologia.
- “monte” - representativo de alguns terrenos não cultivados que por vezes partilhavam confrontações com terrenos da Mesa Arcebispal, não sendo um termo relevante para a ontologia.

A lista de conceitos candidatos extraídos pelo primeiro método apresenta termos iguais aos previamente obtidos, com a exceção de "parte", "partes", "lugar", "prazo" e "dia".

- “parte” - diz respeito à partilha de uma confrontação entre duas terras, p.e. "terreno A parte a norte com terreno B".
- “partes” - plural do termo *parte*, pelo que não é necessário explicitá-lo.
- “lugar” - outro possível tipo de terra ou propriedade.
- “prazo” - representativo do prazo dos pagamentos estipulado num documento legal, redigido pelo escrivão da Mesa Arcebispal e assinado pelo procurador geral do arcebispo ou pelo próprio, no qual se faz o emprazamento da propriedade.
- “dia” - referido no prazo (o documento) a propósito do momento determinado para pagamento do foro e penção à Mesa Arcebispal, sendo também utilizado como parte do termo meio-dia, que nestes fólios é sinónimo de "sul", (partir a meio-dia com um terreno B é o mesmo que partir a sul com um terreno B).

Com a informação apresentada anteriormente, o único termo que foi possível confirmar que é um conceito relevante para a ontologia é o relativo à terra ou propriedade. Porém, este não é o único conceito existente. Após a leitura dos fólios, é possível constatar que existem mais dois conceitos. O primeiro termo que representa um conceito é "título", que representa um grande conjunto de propriedades da Mesa Arcebispal, emprazadas por um pequeno número de emprazadores. O agrupamento destas diferentes propriedades tem como vantagem uma melhor noção geográfica de onde ficam as propriedades. Geralmente, quando um título é referido, também é mencionado o seu nome e o foro que é necessário pagar pelos emprazadores, sendo estas caraterísticas consideradas atributos do título. A razão pela qual este termo não foi detetado por nenhum dos métodos é por se tratar de um termo que não é comum. Como os títulos são agrupamentos vastos de propriedades, muitas vezes só são mencionados uma vez nos fólios para a sua introdução, introdução essa que é seguida da descrição de todas as propriedades pertencentes a esse título. Inclusive, existem fólios que não possuem um título. Nesses casos todas as propriedades pertencem ao último título introduzido nos fólios anteriores. O outro conceito que é possível reconhecer da explicação anterior é o conceito de *emprazador*. Este designa uma pessoa, a quem a Mesa Arcebispal empraza uma ou mais terras com o propósito de as cultivar para subsistir, tendo de pagar um foro à Mesa em si. Muito raramente as propriedades possuem o nome do emprazador, pelo que estabelecer uma relação direta entre o emprazador e as suas terras não forneceria resultados muito bons. Porém, a ligação dos emprazadores com o conceito de título é muito mais fácil de ser efetuada, visto que na introdução do título são sempre listados os emprazadores de propriedades nesse título. A razão pela qual o termo *emprazador* não foi detetado pelos métodos enunciados anteriormente é similar à razão pela qual o termo *título* também não foi extraído, visto que muito raramente é referido o nome do emprazador de um terreno específico.

Outro ponto que é necessário referir é que, apesar da maior parte dos terrenos que produzem alimento serem de cultivo de cereais, nem todos os terrenos desempenham este papel. Se um terreno possuir o tipo "vinha", então a quantidade de vinho produzido não é medida em alqueires, mas sim em almudes, podendo

ainda mais frequentemente ser calculada pelo número de homens necessário para efetuar a sua cava. A razão pela qual não foi extraído nenhum termo relacionado com este aspeto de alguns terrenos é o facto de a maior parte das terras mencionadas não possuírem vinhas. Este ponto faz com que as terras ou propriedades possam ser divididas em dois subconceitos, sendo estes *vinha*, que teria o atributo *cava* e *outros terrenos*, que teria um atributo *semeadura*, de forma a salientar as diferentes caraterísticas que estas terras distintas possuem.

Por último, é possível constatar que não só foram encontrados os conceitos relevantes desta ontologia, como também os termos relativos das relações entre estes e inclusive os atributos destes mesmos conceitos, o que transforma esta etapa na mais relevante de toda a ontologia elaborada. De seguida, será abordada a validação de relações entre elementos da ontologia.

4.2.4 *Validação de Relações entre Elementos da Ontologia*

Após a obtenção dos termos essenciais e dos conceitos, o passo seguinte foi a obtenção de relações entre os diferentes conceitos presentes na ontologia. Porém, a análise dos termos obtidos no final da etapa anterior permitiu ter uma noção não só dos conceitos relevantes para a ontologia a elaborar, mas também de quais as relações existentes entre eles e quais as suas caraterísticas fundamentais. Devido a isto, esta etapa, e principalmente os seus resultados, terá como propósito a validação das relações entre os elementos obtidos na fase anterior, principalmente entre os conceitos e os atributos.

O método utilizado para executar esta etapa da ontologia foi o algoritmo de regras de associação utilizando o algoritmo *Apriori*, referido anteriormente na secção 3 desta dissertação. Para calcular quão forte é uma relação entre dois termos foi utilizada a função de *lift*, que mede a qualidade e grau de interesse na relação, combinando as funções de suporte e confiança (Harun et al. 2017).

Descrição do método

Primeiramente, como nos outros passos da ontologia, foi necessário obter o conteúdo dos fólios; conteúdo esse que foi colocado num *array*, sendo cada elemento uma frase dos fólios. Após este passo foram removidos todas os elementos frásicos que não contivessem carateres alfa numéricos, como, por exemplo, vírgulas. De seguida, utilizando a biblioteca *Pandas* (*Pandas Library*), os *outputs* da etapa anterior foram importados, sendo prontamente descartadas as palavras independentes do contexto, mencionadas na etapa anterior, nomeadamente, "elle", "hum", entre outras. De seguida, foi aplicado o algoritmo *Apriori* a todo o conteúdo dos fólios, filtrado das maneiras já referidas, armazenando num dicionário as relações estabelecidas entre diferentes termos, desde que estes estivessem presentes nos *outputs* da etapa anterior. Só foram consideradas as relações nas quais os termos apareciam na mesma frase um mínimo de 20 vezes. Após a aplicação deste filtro foram definidas as funções de suporte, confiança e *lift*. Os resultados foram armazenados num outro dicionário. Os pares de termos foram depois reordenados neste dicionário com base no valor da sua *lift*, desde os pares com maior valor de *lift* e, por conseguinte, mais fortes, até aos pares menos fortes. Os resultados desta etapa encontram-se apresentados no anexo E1.

Resultados obtidos (relações com lift acima de 1.5)

Observando os resultados obtidos, à medida que se desce na lista, consegue-se perceber o decréscimo na força das relações entre os termos, sendo as últimas relações não representativas de relações diretas entre termos. Estes resultados são discutidos com maior detalhe na secção seguinte.

Análise dos resultados

A análise de resultados focou-se nos pares de termos que possuem um *lift* superior a 2.0, visto que ao analisar as relações com um *lift* abaixo deste, é notório que não retratam relações diretas, ou seja, não são instâncias que representem conceitos ou atributos que estejam diretamente relacionados. Um exemplo deste caso é a relação entre os termos *semeadura* e *varas* que possui um *lift* de 1.7428400383141762. O primeiro termo é representativo do cultivo de um terreno, enquanto o segundo representa a unidade com que as dimensões dos terrenos eram medidas nos fólios. Estes termos, apesar de representarem dois atributos do mesmo conceito e, por isso, coexistem normalmente nas mesmas frases nos fólios, não possuem uma ligação direta, não tendo esta relação muito valor para o contexto da ontologia.

No anexo E2 encontra-se a análise efetuada das relações com mais de 2.0 de *lift*.

Esta etapa serviu para confirmar e validar alguns pontos relativos à ontologia e algumas das conclusões a que se tinha chegado na etapa prévia, maioritariamente ligados às relações existentes entre conceitos e os seus atributos. De seguida, abordaremos como foram obtidas as instâncias desta ontologia.

4.2.5 *Obtenção de instâncias*

Tendo já obtido nos passos anteriores os conceitos, as relações e os atributos, à volta dos quais é possível estruturar uma ontologia, o passo seguinte foi a obtenção de instâncias. A abordagem efetuada deteta instâncias com base em padrões que são repetidos ao longo dos fólios.

Esta abordagem depende da informação recolhida previamente, principalmente na etapa da formação de conceitos. Este facto deve-se ao uso de palavras pertencentes à lista obtida no final dessa etapa para detetar padrões existentes no texto. Quando um padrão é procurado utilizando um termo não presente nessa lista, é prontamente justificado o uso desse termo. De seguida são explicitados os métodos desta abordagem.

Descrição da abordagem

O conteúdo existente nesta abordagem foi separado em duas componentes: uma que obtém informação relativa às propriedades em si e as relaciona com os respetivos títulos e outra que obtém informação dos títulos, tentando relacioná-los com os emprazadores. O primeiro componente a ser discutido é o que obtém informação das propriedades em particular. Este componente foi divido em três fases, a inicial, a intermédia e a final, com base na quantidade de informação já extraída acerca da propriedade.

Primeiramente foi necessário ler o conteúdo dos fólios, com funções elaboradas propositadamente para esse efeito, e depois realizar as operações de filtragem do texto, tais como a remoção de cardinais, das notas existentes entre parênteses retos ao longo dos fólios e a remoção de pontos finais aleatórios, entre outros.

Primeira Componente: Fase Inicial

Com o intuito de se conseguir depois separar as propriedades pelos respetivos títulos para facilitar a descoberta de instâncias das relações entre estes dois conceitos, foi necessário separar os títulos, bem como separar a informação relativa ao título das propriedades desse mesmo título. De seguida, começou a ser extraída informação acerca dos fólios começando com os possíveis tipos de terrenos que podem ser associados a cada terra. Foi observado que, normalmente, aquando da utilização da palavra "outro(a)", a palavra seguinte indica o tipo do terreno que está a ser descrito. Após a obtenção dos tipos é realizada a obtenção das confrontações das propriedades. Como foi descrito no passo de formação de conceitos, o termo "parte" representa o estabelecimento da confrontação entre a propriedade a ser descrita atualmente e uma propriedade vizinha, sendo esta uma das relações mais comuns nos fólios, pelo que este termo foi utilizado para determinar as confrontações das propriedades em si. É ainda de notar que nos fólios a expressão "pella parte" não faz referência à terra com que faz confrontação, mas sim à dimensão da confrontação em si, sendo sinónima de "pello lado", "pella banda", o que, consequentemente, obriga a que seja verificado que a palavra anterior a "parte" nos fólios não seja "pella". Mais raramente é utilizado no lugar de "parte" o termo "confronta", pelo que o aparecimento deste termo também foi considerado. A descrição de pequenos conjuntos de propriedades, e às vezes propriedades singulares, é descrita muitas vezes pela palavra "Ittem" ou por algumas variações ("Item", It., H.), sendo que nas raras vezes em que este termo não é utilizado para efetuar esta separação, é utilizado um ponto final. Por conseguinte, estes separadores foram utilizados para dividir as diferentes instâncias de propriedades. Para concluir esta fase, foi verificado se a propriedade pertence a algum dos títulos obtidos previamente, estabelecendo a ligação entre as duas instâncias, se for esse o caso.

Um exemplo de uma instância dos resultados obtidos nesta fase pode ser consultado no Anexo F1.

Este *output* é composto por quatro elementos. O primeiro menciona o conjunto seguido de fólios em que esta propriedade se encontra. O segundo indica a informação relativa à propriedade após a descrição da confrontação (a seguir à utilização de um termo igual ou com o mesmo significado que "parte"). O terceiro indica a primeira secção da frase que normalmente introduz a propriedade. O quarto e último elemento é o título a que a propriedade pertence. Este é mantido na íntegra para auxiliar melhor na interligação dos *outputs* finais deste componente com o componente que extrai informação relativa aos títulos.

Primeira Componente: Fase Intermédia

Nesta secção, a primeira ação a ser efetuada foi verificar quais as diferentes confrontações que a propriedade atual tem. Existem quatro possibilidades diferentes, que foram termos extraídos na etapa de formação de conceitos: "norte", "sul", "nascente" e "poente". Porém, estes termos nem sempre são os termos utilizados. Outros termos utilizados são "aguião", no lugar de "norte", "vendaval" e "meio-dia", no lugar de "sul", e "levante", no lugar de poente. Nesta verificação foram também extraídas quaisquer informações acerca de

confrontações de outras orientações do excerto relativo à confrontação que se estava a analisar, verificando ao mesmo tempo se se encontrava presente a expressão "e das demais partes", pois a existência desta expressão indica que todas as confrontações de outras orientações que não foram indicadas são relativas à mesma propriedade vizinha. Estas verificações foram efetuadas para todas as orientações, visto que a ordem em que as orientações das confrontações são mencionadas é variável. Para todas as confrontações, caso estas sejam a última, também foi verificado se existe alguma informação acerca do possível cultivo da terra. Visto que a expressão utilizada para introduzir o tema do cultivo da terra é sempre "levara de semeadura", esta foi inicialmente utilizada para efetuar a divisão entre a última confrontação e o cultivo. Porém, foi verificado que nem todos os terrenos inventariados oferecem esta expressão, em particular as vinhas. No que toca a este tipo de terreno, em vez de ser indicada a quantidade de cereal, em alqueires, necessária para o semear, é indicado quantos homens são necessários para efetuar a cava, sendo a expressão utilizada neste cenário "levara de cava". Esta situação fez com que, em vez de utilizar a expressão "levara de semeadura" para efetuar a separação entre a secção relativa ao cultivo e a secção relativa à confrontação, fosse utilizada somente a palavra "levara", em todas as variações em que esta é escrita. Após a verificação do cultivo foi também verificado se existe alguma informação relativa às dimensões do terreno em si. Para obter esta informação simplesmente é verificado se no excerto da confrontação se encontra o termo "largo" ou o termo "comprido", termos estes já explicitados na etapa de formação de conceitos. Embora estes últimos atributos da propriedade se encontrem normalmente depois de terem sido mencionadas as diferentes confrontações dessa mesma propriedade, existem alguns casos em que são mencionados antes, pelo que as dimensões e cultivo também são pesquisadas na primeira secção da frase que tinha também sido armazenada na fase anterior. A última informação determinada nesta etapa é relativa à obtenção do nome da propriedade, que em alguns casos é apresentado, particularmente nos fólios iniciais. Os padrões que foram explorados na procura dos nomes dos terrenos são:

- nome1 de nome2: alguns exemplos são "Veiga do Sobredo" e "Vinha da Lenta";
- tipo chamado de nome: alguns exemplos são "Campo chamado de Pumar da Froỳa" e "Campo chamado da Veiga";
- tipo que se chama de nome: alguns exemplos são "Campo que se chama Esquenta cabeça" e "Campo que se chama linhares de Syma";
- tipo a que chamam de nome: um exemplo é "cortinha a que chamam do Ameeiral".

Um exemplo de uma instância dos resultados obtidos nesta fase pode ser consultado no Anexo F2.

O exemplo é composto por oito elementos. Como no *output* anterior, o primeiro elemento faz menção à sequência de fólios a que esta terra pertence. O segundo elemento é representativo da primeira parte da frase antes da menção das suas confrontações. De salientar que este é um dos casos em que o atributo nome do terreno não se encontra presente nesta secção da frase. O terceiro elemento contém informação acerca das dimensões do terreno, quer comprimento, quer largura. O quarto elemento possui informação acerca do cultivo do terreno. O quinto elemento contém informação relativa a todas as confrontações desta terra. Algo a salientar é que a confrontação norte é a mesma que as confrontações nascente e sul, devido ao uso da expressão "das

demais partes", como foi explicitado anteriormente. O sexto elemento é a totalidade da frase relativa a este terreno. O sétimo elemento seria o nome da propriedade. Como a informação sobre o nome do terreno não existe neste exemplo, este elemento simplesmente funciona como descrição do tipo de terreno. O último elemento é o título a que a propriedade pertence, na íntegra, tal como na fase anterior.

Primeira Componente: Fase Final

Esta fase tem apenas como propósito a preparação da informação que foi obtida anteriormente para exportação para um ficheiro CSV. A principal alteração envolveu transformar os dicionários existentes dentro de cada elemento correspondente a uma terra, que são relativos às dimensões e confrontações, em simples elementos desse mesmo *array*. Inicialmente a extração dos emprazadores era efetuada nesta secção. Porém, como referido anteriormente, a existência de informação sobre os emprazadores nas propriedades é muito raro nestes fólios, pelo que a procura por esta informação passou a ser efetuada pelos títulos, visto que estes usualmente incluem os nomes dos emprazadores das suas propriedades. Um exemplo de uma instância dos resultados obtidos nesta fase pode ser consultado no Anexo F3.

O *output* final é composto por doze elementos. As únicas alterações comparativamente ao *output* anterior envolveram a adição do terceiro campo indicativo do tipo de terreno e a separação dos elementos dos objetos relativos às dimensões e confrontações em elementos distintos da propriedade. Esta informação é depois armazenada num ficheiro CSV próprio.

Segunda Componente

A segunda componente possui um início igual ao do anterior, na medida em que o conteúdo dos fólios é recolhido e filtrado como na outra componente. Também é efetuada a separação do conteúdo dos fólios por títulos como na outra componente, embora neste caso a informação relativa às propriedades seja descartada, visto que para esta componente só tem relevância a informação relativa aos títulos presentes nos fólios. Todas as propriedades de um título vêm depois da expressão "tem as propriedades", sendo por isso esta expressão utilizada como separador. Existem três informações que é possível extrair dos títulos, que são o nome do título, os emprazadores deste título e o foro a pagar à Mesa Arcebispal por parte destes. Para obter informação acerca do foro foram utilizadas as palavras "paga" ou "foro" como separadores e para obter informação dos emprazadores foi utilizada a palavra "possui". Um exemplo de uma instância dos resultados obtidos por este componente pode ser consultado no Anexo H.

O *output* desta componente possui quatro elementos. O primeiro diz respeito a toda a informação existente na descrição do título. O segundo indica qual o nome do título em questão. O terceiro elemento contém os emprazadores do título em questão. Por fim, o quarto e último elemento contém informação acerca do foro a ser pago pelos emprazadores à Mesa Arcebispal.

Desta maneira concluímos a aplicação da metodologia utilizada para a obtenção da ontologia. De referir que não foi acrescentada nenhuma etapa relativa à deteção de axiomas no texto, uma vez que é preferível detetá-los manualmente, sendo consequentemente encontrado um axioma, apresentado na secção 5.1.

5

A ONTOLOGIA FINAL

5.1 ESTRUTURA DA ONTOLOGIA

A estrutura da ontologia que foi desenvolvida acha-se apresentada na Figura 17, integrando na sua constituição conceitos, relações taxonómicas e não taxonómicas, atributos e um axioma (definido manualmente).

Os conceitos encontram-se representados por quadrados, enquanto as relações taxonómicas estão representadas por setas mais grossas do que as utlizadas para representar relações entre os conceitos e os seus atributos, o que as permite distinguir das relações não taxonómicas por terem o verbo *ser* nas suas descrições.

Nesta ontologia, existem três tipos de atributos distintos, nomeadamente: simples, que são compostos por um único valor (*e.g.* Preço), multivariado, que pode possuir mais do que um valor diferente (*e.g.* "Tipo"), e compostos, que possuem na sua constituição mais do que um atributo simples (*e.g.* Dimensões). Os dois primeiros tipos são representados por círculos, enquanto os atributos compostos são representados por elipses. Além disto, também foi definido, de forma manual, um axioma, que pode ser observado na parte inferior da figura referida anteriormente. Os elementos da ontologia, excetuando o axioma, bem como os seus tipos, encontram-se apresentados no anexo H.

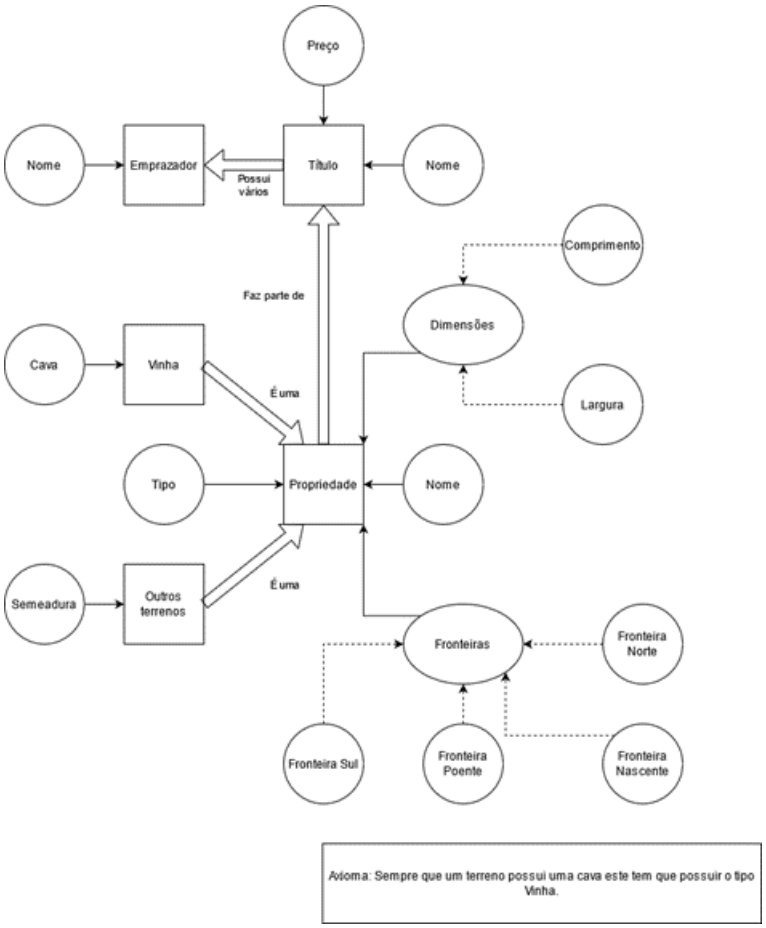

Figura 17: Estrutura da ontologia final

5.2 EXEMPLOS DE INTERAÇÕES ENTRE INSTÂNCIAS DOS ELEMENTOS

5.2.1 *Relação entre título e emprazadores*

Um exemplo da interação entre emprazadores e um título, bem como das relações entre um título e os seus atributos, podem ser identificados através do seguinte título:

"Titulo do Cazal do Gial, que trouxe Gildo Gial, Sito na / freyguezia de São Martinho d'Aluarenga digo de Aluaredo, / que possuem Francisco Alres' e Affonso Vaz', e Ine alres', Mulher não Cazada, Ruy Gonçalues filho que ficou de João Artur e / os mais herdejros do dito João Artur não mostrarão prazo e dicerão que não sabião que nunqª o ouuesse, pagasse de foro deste casal ao snor Arçº e sua Camera de São Payo en cada hu' anno / quatro alqueires de Pão meado milho e Centeo, e hua' lamprea, e as propriedades são as"

Este sistema, (em particular da deteção de padrões), aquando da aplicação a este título, permitiu-nos concluir que:

- **Nome e descrição:** "Titulo do Cazal do Gial, que trouxe Gildo Gial, Sito na / freyguezia de São Martinho d'Aluarenga digo de Aluaredo, / que ",
- **Emprazadores deste título:** "m Francisco Alres' e Affonso Vaz', e Ine alres', Mulher não Cazada, Ruy Gonçalues filho que ficou de João Artur e / os mais herdejros do dito João Artur não mostrarão prazo e dicerão que não sabião que nunqª o ouuesse, pagasse ",
- **Preço do foro a pagar à Mesa Arcebispal:** "deste casal ao snor Arçº e sua Camera de São Payo en cada hu' anno / quatro alqueires de Pão meado milho e Centeo, e hua' lamprea, "

Estas informações permitem concluir que existe uma relação entre:

- a instância da classe Título e uma instância do atributo nome com o valor "Titulo do Cazal do Gial"
- a instância da classe Título e uma instância do atributo preço com o valor "quatro alqueires de Pão meado milho e Centeo, e hua' lamprea"
- a instância de Título e quatro instâncias da classe Emprazador, que é não taxonómica com o valor "Tem como emprazador"
- cada uma das instâncias da classe Emprazador e as respetivas instâncias do atributo nome, que possuem os valores "Francisco Alres", "Affonso Vaz", "Ine alres" e "Ruy Gonçalues"

Estas relações estão representadas na Figura 18.

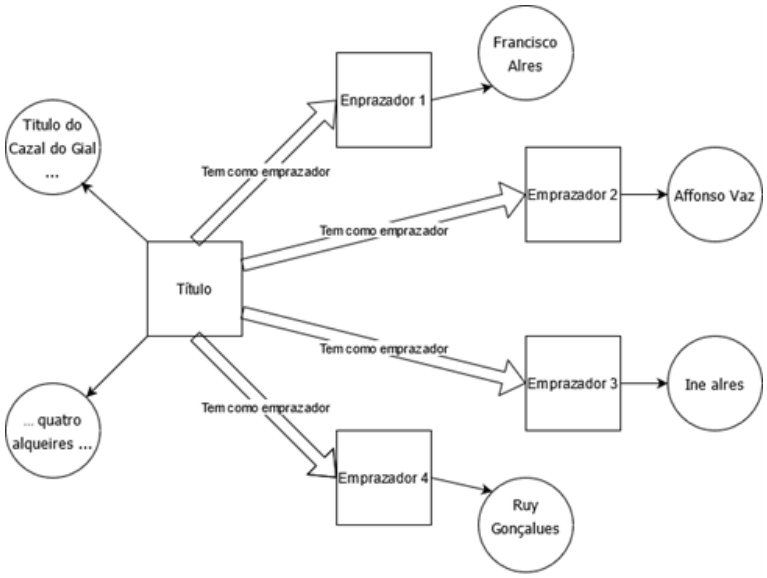

Figura 18: Exemplo da relação entre emprazadores e um título

5.2.2 *Relações taxonómica e não taxonómica envolvendo uma propriedade*

Um exemplo das relações taxonómicas e não taxonómicas envolvendo uma propriedade é a da seguinte propriedade:

"O campo do Ricoy que traz / Ruy Gonçalues tem de comprido de Nacente a poente sesenta e quatro varas, e de largo de / Norte a sul trinta e duas, leuara de semeadura tres alqueires de centeo, pouco mais ou menos / parte do Nacente com a estrada publica, que uay pera Melgaco e Valladares, e outras partes, e do Poente co' terras de São Martinho e de Paderne, que traz o Mesmo Ruy Gonçalues / e João Roiz do souto, e do Norte tãobem com os Mesmos e do Vendaual co' a Mesma estrada / acima declarada."

A aplicação do sistema que desenvolvemos sobre esta propriedade, em particular a parte da deteção de padrões, permitiu concluir que:

- **Nome da propriedade:** campo do Ricoy,
- **Semeadura:** :"['de semeadura tres alqueires de centeo, pouco mais ou menos']",
- **Título a que propriedade pertence:** "Titulo do Cazal do Gial, que trouxe Gildo Gial, Sito na / freyguezia de São Martinho d'Aluarenga digo de Aluaredo, / que possuem Francisco Alres' e Affonso Vaz', e Ine alres', Mulher não Cazada, Ruy Gonçalues filho que ficou de João Artur e / os mais herdejros do dito João Artur não mostrarão prazo e dicerão que não sabião que nunqª o ouuesse, pagasse de foro deste casal ao snor Arçº e sua Camera de São Payo en cada hu' anno / quatro alqueires de Pão meado milho e Centeo, e hua' lamprea, e as propriedades são as"

A informação apresentada no excerto anterior permite concluir que existe uma relação entre:

- a instância da classe Título e a instância da classe Propriedade, não taxonómica, com o valor de "Faz parte de"
- a instância da classe Propriedade e a instância do atributo nome com o valor "campo do Ricoy"
- a instância da classe Propriedade e a instância da subclasse Outro Terreno (visto que o terreno não é uma vinha), taxonómica com o valor "É uma"
- a instância da classe Outro Terreno e a instância do atributo semeadura com o valor "de semeadura tres alqueires de centeo, pouco mais ou menos"

As interações entre os diversos elementos da ontologia relativos a este elemento podem ser observadas na Figura 19.

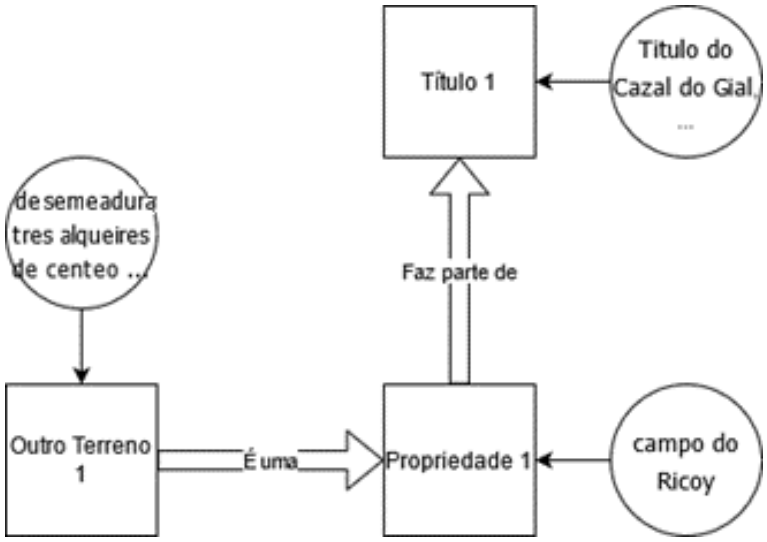

Figura 19: Exemplo das relações de uma propriedade

5.3 EXEMPLO DE UMA PROPRIEDADE E DOS SEUS ATRIBUTOS

Além dos resultados apresentados na secção anterior, a aplicação do sistema, em particular da sua componente de deteção de padrões, permitiu, também, concluir que:

- **Nome da propriedade:** campo do Ricoy,
- **Tipo(s) da propriedade:** campo,
- **Dimensões:**
  - **Comprimento:** “de Nacente a poente sesenta e quatro varas”,
  - **Largura:** "de Norte a sul trinta e duas,",
- **Fronteiras:**
  - **Fronteira Norte:** “tãobem com os Mesmos,co’ a Mesma estrada acima declarada.”,
  - **Fronteira Poente:** "co’ terras de São Martinho e de Paderne, que traz o Mesmo Ruy Gonçalues e João Roiz do souto,",
  - **Fronteira Nascente:** "com a estrada publica, que uay pera Melgaco e Valladares, e outras partes,"

É possível perceber, através do excerto anterior, que existe uma relação entre:

- a instância da classe Propriedade e a instância do atributo simples "tipo"com o valor “campo”
- a instância da classe Propriedade e a instância do atributo simples "nome"com o valor “campo do Ricoy”

- a instância da classe Propriedade e a instância do atributo composto "dimensões", que por sua vez se divide em instâncias dos atributos simples comprimento e largura, cujos valores são, respetivamente, “sesenta e quatro varas” e “trinta e duas”
- a instância da classe Propriedade e a instância do atributo composto "fronteira", que por sua vez se divide em instâncias dos atributos simples fronteira norte, fronteira poente e fronteira sul, cujos valores são, respetivamente, “co’ a Mesma estrada acima declarada”, “co’ terras de São Martinho e de Paderne” e “com a estrada publica”

As relações existentes entre estes diversos elementos da ontologia encontram-se descritas na figura seguinte.

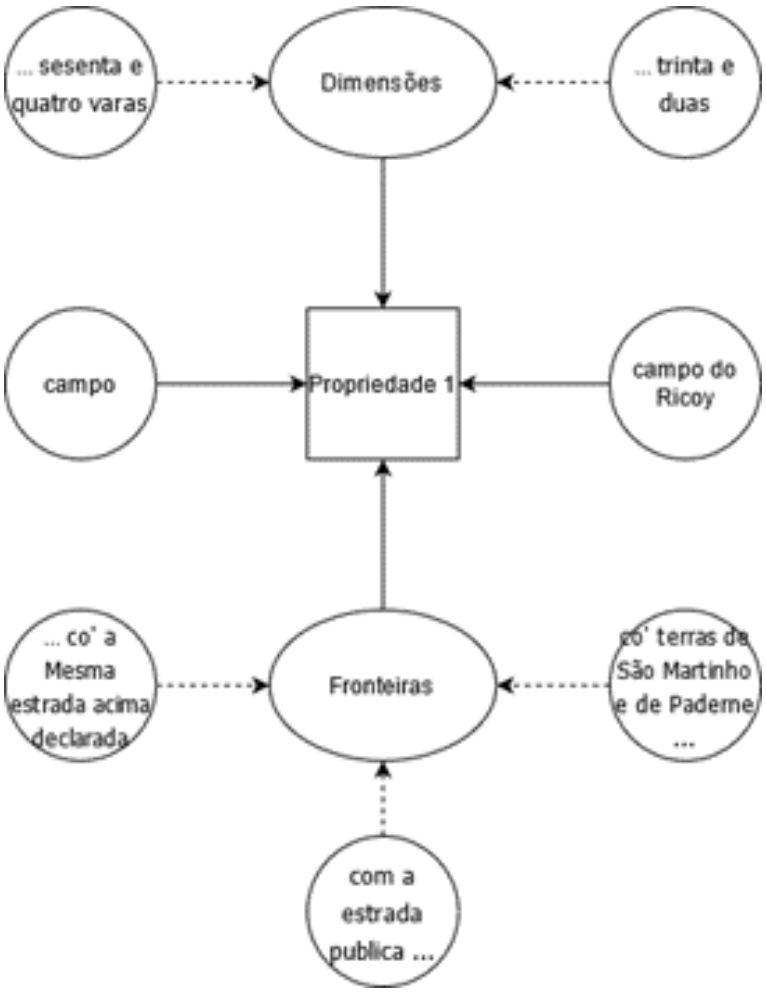

Figura 20: Relação duma propriedade com os seus atributos

6

CONCLUSÕES E TRABALHO FUTURO

6.1 CONCLUSÕES

É inegável o impacto que as ontologias têm na atualidade. Tal pode ser verificado não só pelo número de aplicações distintas que estas possuem em domínios radicalmente diferentes, mas também pela forma excelente como estas permitem, dentro das suas estruturas, armazenar e representar conhecimento. Por ser um tema tão relevante e por terem uma utilização tão vasta e necessária em áreas muito distintas, é necessário relevar quão importante é o processo de elaboração de ontologias, que deve conduzir a um produto (uma ontologia) bem elaborado e bem estruturado. Esta circunstância constribui para garantir a representação correta do conhecimento de diferentes domínios e, acima de tudo, prevenir a divulgação de conhecimento inválido, quer da interpretação errónea do conteúdo ontológico, devido à possível existência de ambiguidades, quer da informação errada que possa estar representada numa ontologia mal elaborada.

De forma a evitar tais representações incorretas da informação é necessário estudar, planear e elaborar um plano bem definido que seja capaz de sustentar o processo de elaboração de uma ontologia e garantir, tanto quanto possível, que é levado a bom termo. Assim, nesta dissertação fizemos uma avaliação muito cuidada de todas as etapas que definimos na nossa metodologia de trabalho, o que implicou a criação de tarefas bem concretas de validação e de análise de resultados em todas as etapas do sistema criado. O nível de exigência e correção requeridos para cada uma das etapas foi alto, devido a estarmos envolvidos num processo (semi)automático de construção de uma ontologia, requerendo, por isso, um maior controlo dos resultados do que seria necessário se estivéssemos envolvidos num processo de construção manual de uma ontologia.

O Livro das Propriedades é constituído por centenas de fólios, que, por sua vez, contêm muita informação acerca de diversos títulos do território pertencente ao Arcebispado de Braga e das inúmeras propriedades contidas dentro desses títulos. A quantidade de informação contida dentro deste códice, por conseguinte, possui grandes dimensões, sendo a sua divisão um grande passo em direção à sua mais cabal compreensão.

A estrutura da ontologia obtida com este trabalho de dissertação permite fazer a divisão dos diversos elementos constituintes dos fólios, de forma que os diferentes aspetos das propriedades sejam separados, permitindo uma consulta mais fácil da informação neles armazenada. Isto deve-se à facilidade em encontrar a informação pretendida de uma propriedade ou outra entidade, e à rapidez dessa procura, devido a esta não ser efetuada em fólios densos, mas sim em instâncias organizadas, como consequência do trabalho realizado nesta

dissertação. Para além destes pontos, também é possível agora analisar as caraterísticas das diversas entidades presentes nos fólios de forma totalmente separada, uma consequência direta da divisão da informação em diversos setores. Um exemplo disso é que, após a realização da ontologia, com base na extração e segregação da informação relativa às dimensões das propriedades, foi calculado um valor aproximado da área total de terreno que a Mesa Arcebispal possuía aquando da escrita dos fólios. Porém, não se conseguiu determinar toda a área de forma conclusiva, devido a não terem sido registadas no Livro as dimensões de alguns dos terrenos referidos nos fólios. Além disso, também foi possível determinar a quantidade de cereal que era possível produzir nos terrenos e quantos homens seriam necessários para efetuar a cava em todas as vinhas mencionadas. Conseguiu-se assim, de forma semiautomática, descobrir dois elementos bastante preciosos para qualquer processo de estudo do Livro das Propriedades.

Em suma, pode-se concluir que o sistema de elaboração da ontologia criado foi um processo que, apesar de apresentar algumas dificuldades, foi bem-sucedido no seu objetivo fundamental: a construção e validação de um ontologia bem estruturada e representativa do universo representado nos fólios do Livro das Propriedades.

6.2 TRABALHO FUTURO

A informação que está contida nos fólios do Livro das Propriedades encontra-se representada de forma adequada na ontologia construída. Para além de alguns filtros que podem ser colocados para reduzir a quantidade de texto de um ou de outro elemento, pensamos não ser necessário fazer mais alterações ao conteúdo da ontologia estabelecida. Logo, a determinação de eventuais linhas de trabalho futuro envolverá, maioritariamente, o aproveitamento da ontologia em si e do seu conteúdo, mais do que o seu melhoramento.

No entanto, poderão no futuro ser testadas outras técnicas ou utilizadas outras metodologias, com o objetivo de obter uma ontologia mais completa.

A ontologia desenvolvida será integrada numa aplicação já existente que foi criada especificamente para fazer o acolhimento e a análise do conteúdo dos fólios do Livro das Propriedades. Nessa aplicação poderão ser efetuadas pesquisas sobre os seus diversos elementos constituintes e, consequentemente, obter um leque de conhecimento mais vasto acerca dos fólios. Por conseguinte, seria interessante desenvolver uma *API* de consulta da ontologia que, através de uma *interface* específica, permitirá o acesso ao conhecimento incorporado na ontologia agora desenvolvida. Com a implementação desta *API* e *interface*, poder-se-á usufruir em pleno da estrutura e do conteúdo armazenado na ontologia de uma forma bastante simplificada e amigável.

BIBLIOGRAFIA

Albawi, Saad, Tareq Abed Mohammed e Saad Al-Zawi. 2018. "Understanding of a convolutional neural network". Em *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017.* ISBN: 9781538619490. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

Alom, Md Zahangir, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A.S. Awwal e Vijayan K. Asari. 2019. *A state-of-the-art survey on deep learning theory and architectures.* https://doi.org/10.3390/electronics8030292.

Asim, Muhammad Nabeel, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood e Hafiza Mahnoor Abbasi. 2018. *A survey of ontology learning techniques and applications.* https://doi.org/10.1093/database/bay101.

Al-Aswadi, Fatima N., Huah Yong Chan e Keng Hoon Gan. 2020. "Automatic ontology construction from text: a review from shallow to deep learning trend". *Artificial Intelligence Review,* ISSN: 15737462. https://doi.org/10.1007/s10462-019-09782-9.

Barros, Anabela. 2018. "Das Palavras de que os Dicionários não Rezam: Um Dicionário Inédito da Língua Portuguesa".

———. Março de 2019. "Apontamentos lexicais sobre o Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga: designações de terras e outros aspetos das propriedades", 393–428. ISBN: 978-989-26-1755-8 / Digital ISBN 978-989-26-1756-5. https://doi.org/10.14195/978-989-26-1756-5.

Bluteau, Raphael. 1712-28. "Vocabulario Portuguez e Latino".

Boytcheva, Svetla. Janeiro de 2002. "Overview of Inductive Logic Programming (ILP) Systems" ().

Buitelaar, Paul, Philipp Cimiano e Bernardo Magnini. 2004. "Ontology Learning from Text : An Overview". *Learning.*

Campos, Ricardo, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes e Adam Jatowt. 2020. "YAKE! Keyword extraction from single documents using multiple local features". *Information Sciences,* ISSN: 00200255. https://doi.org/10.1016/j.ins.2019.09.013.

Campos, Ricardo, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes e Adam Jatowt. 2018. "A text feature based automatic keyword extraction method for single documents". Em *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* ISBN: 9783319769400. https://doi.org/10.1007/978-3-319-76941-7_63.

Dicionário Online de Português. http://www.lexico.pt. Accessed: 2021-11-08.

Fan, Shuchuan, Li Zhang e Zian Sun. 2009. "An ontology based method for business process integration". Em *Proceedings - 2009 International Conference on Interoperability for Enterprise Software and Applications, IESA 2009.* ISBN: 9780769536521. https://doi.org/10.1109/I-ESA.2009.31.

Faure, David, Claire Nedellec e Celine Rouveirol. 1998. "Acquisition of semantic knowledge using machine learning methods: The system ASIUM". *Laboratoire de Recherche en Informatique.*

Gennari, John H., Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubezy, Henrik Eriksson, Natalya F. Noy e Samson W. Tu. 2003. "The evolution of Protégé: An environment for knowledge-based systems development". *International Journal of Human Computer Studies,* ISSN: 10715819. https://doi.org/10.1016/S1071-5819(02)00127-1.

Girardi, Rosario. 2010. "Guiding ontology learning and population by knowledge system goals". Em *KEOD 2010 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development.* ISBN: 9789898425294. https://doi.org/10.5220/0003119404800484.

Gruber, Thomas R. 1995. "Toward principles for the design of ontologies used for knowledge sharing". *International Journal of Human - Computer Studies,* ISSN: 10959300. https://doi.org/10.1006/ijhc.1995.1081.

Harun, Nur Ashikin, Mokhairi Makhtar, Azwa Abd Aziz, Zahrahtul Amani Zakaria, Fadzli Syed Abdullah e Julaily Aida Jusoh. 2017. "The application of Apriori algorithm in predicting flood areas". *International Journal on Advanced Science, Engineering and Information Technology,* ISSN: 24606952. https://doi.org/10.18517/ijaseit.7.3.1463.

Hassan, Abdalraouf, e Ausif Mahmood. 2018. "Convolutional Recurrent Deep Learning Model for Sentence Classification". *IEEE Access,* ISSN: 21693536. https://doi.org/10.1109/ACCESS.2018.2814818.

Hejl, Leo. 2014. "Evolution of the Conception of Parts of Speech".

Huang, Maojun. 2010. "On the concept of geographic ontology - From the viewpoints of philosophy ontology, information ontology and spatial ontology". Em *2010 18th International Conference on Geoinformatics, Geoinformatics 2010.* ISBN: 9781424473021. https://doi.org/10.1109/GEOINFORMATICS.2010.5567723.

Jiang, Xing, e Ah Hwee Tan. 2005. “Mining ontological knowledge from domain-specific text documents”. Em *Proceedings - IEEE International Conference on Data Mining, ICDM*. ISBN: 0769522785. https://doi.org/10.1109/ICDM.2005.97.

Jurafsky, Daniel, e James H Martin. 2008. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (second edition)”. *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition,* ISSN: 08912017.

Kim, Jong Woo, Jordi Conesa Caralt e Julia K. Hilliard. 2007. “Pruning bio-ontologies”. Em *Proceedings of the Annual Hawaii International Conference on System Sciences*. ISBN: 0769527558. https://doi.org/10.1109/HICSS.2007.455.

Kim, Yoon. 2014. “Convolutional neural networks for sentence classification”. Em *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. ISBN: 9781937284961. https://doi.org/10.3115/v1/d14-1181. arXiv: 1408.5882.

Kiong, Yip Chi, Sellappan Palaniappan e Nor Adnan Yahaya. 2011. “Health ontology system”. Em *2011 7th International Conference on Information Technology in Asia: Emerging Convergences and Singularity of Forms - Proceedings of CITA’11*. ISBN: 9781612841304. https://doi.org/10.1109/CITA.2011.5999506.

Kurematsu, Masaki, Takamasa Iwade, Naomi Nakaya e Takahira Yamaguchi. 2004. “Doddle II: A domain ontology development environment using a MRD and text corpus”. Em *IEICE Transactions on Information and Systems*.

Larkey, Leah S., Lisa Ballesteros e Margaret E. Connell. 2002. “Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis”. Em *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*.

Le, Quoc V. 2015. “A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks”.

Lecun, Yann, Yoshua Bengio e Geoffrey Hinton. 2015. *Deep learning*. https://doi.org/10.1038/nature14539.

Liu, Weibo, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu e Fuad E. Alsaadi. 2017. “A survey of deep neural network architectures and their applications”. *Neurocomputing,* ISSN: 18728286. https://doi.org/10.1016/j.neucom.2016.12.038.

Maedche, Alexander, e Steffen Staab. 2000. “Discovering Conceptual Relations from Text”. *Frontiers in artificial intelligence and applications ; 54*.

Mahgoub, Hany. 2006. “Mining Association Rules from Unstructured Documents”. *International Journal of Applied Mathematics and Computer Sciences Volume.*

Merriam-Webster’sAxiom. https://www.merriam-webster.com/dictionary/axiom. Accessed: 2021-10-26.

Merriam-Webster’sOntology. https://www.merriam-webster.com/dictionary/ontology. Accessed: 2021-10-26.

Najafabadi, Maryam M., Flavio Villanustre, Taghi M. Khoshgoftaar, Naeem Seliya, Randall Wald e Edin Muharemagic. 2015. “Deep learning applications and challenges in big data analytics”. *Journal of Big Data,* ISSN: 21961115. https://doi.org/10.1186/s40537-014-0007-7.

Njike-Fotzo, H, e P Gallinari. 2004. “Learning Generalization/Specialization Relations between Concepts–Application for Automatically Building Thematic Document Hierarchies”. Em *RIAO 2004.*

NLTK Library. https://www.nltk.org/howto/portuguese_en.html. Accessed: 2021-11-08.

Noy, Natalya F., e Deborah L. McGuinness. 2001. *Ontology Development 101: A Guide to Creating Your First Ontology.* Relatório técnico. https://doi.org/10.1016/j.artmed.2004.01.014.

OS Library. https://docs.python.org/3/library/os.html. Accessed: 2021-11-08.

OS.Path Library. https://docs.python.org/3/library/os.path.html. Accessed: 2021-11-08.

Pandas Library. https://pandas.pydata.org. Accessed: 2021-11-08.

Petrucci, Giulio, Chiara Ghidini e Marco Rospocher. 2016. “Ontology learning in the deep”. Em *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* ISBN: 9783319490038. https://doi.org/10.1007/978-3-319-49004-5_31.

Poli, Roberto, Michael Healy e Achilles Kameas. 2010. *Theory and applications of ontology: Computer aplications.* ISBN: 9789048188468. https://doi.org/10.1007/978-90-481-8847-5.

Polyglot. https://polyglot.readthedocs.io/en/latest/. Accessed: 2021-12-20.

Protégé. http://protege.stanford.edu. Accessed: 2021-10-26.

Rees, Reinout Van. 2003. “Clarity in the usage of the terms ontology, taxonomy an classification”. *CIB REPORT.*

Requests Library. https://docs.python-requests.org/en/latest/. Accessed: 2021-11-09.

Richardson, Leonard. 2016. “Beautiful Soup Documentation”. *media.readthedocs.org.*

Santos, Cláudia da Silva Amaral. 2010. *Terminologia e Ontologias: Metodologias para Representação do Conhecimento.*

Serra, Ivo, Rosario Girardi e Paulo Novais. 2014. "Evaluating techniques for learning non-taxonomic relationships of ontologies from text". *Expert Systems with Applications,* ISSN: 09574174. https://doi.org/10.1016/j.eswa.2014.02.042.

Shamsfard, Mehrnoush. 2006. "Learning concepts, taxonomic and non-taxonomic relations from texts". Em *IEEE Intelligent Systems.* ISBN: 1424401968. https://doi.org/10.1109/IS.2006.348404.

Shao, Huajie Z., Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang e Tarek Abdelzaher. 2020. "ControlVAE: Controllable variational autoencoder". Em *37th International Conference on Machine Learning, ICML 2020.* ISBN: 9781713821120. arXiv: 2004.05988.

Shen, Yelong, Xiaodong He, Jianfeng Gao, Li Deng e Grégoire Mesnil. 2014. "Learning semantic representations using convolutional neural networks for web search". Em *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web.* ISBN: 9781450327459. https://doi.org/10.1145/2567948.2577348.

Song, Qiuxia, Jin Liu, Xiaofeng Wang e Jin Wang. 2014. "A novel automatic ontology construction method based on web data". Em *Proceedings - 2014 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2014.* ISBN: 9781479953905. https://doi.org/10.1109/IIH-MSP.2014.194.

Tran, Oanh Thi, Cuong Anh Le, Thuy Quang Ha e Quynh Hoang Le. 2009. "An experimental study on Vietnamese POS tagging". Em *2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009.* ISBN: 9780769539041. https://doi.org/10.1109/IALP.2009.14.

Vazifedoost, A. R., F. Oroumchian e M. Rahgozar. 2007. "Finding similarity relations in presence of taxonomic relations in ontology learning systems". Em *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2007.* ISBN: 1424407052. https://doi.org/10.1109/CIDM.2007.368875.

Wong, Wilson, Wei Liu e Mohammed Bennamoun. 2012. "Ontology learning from text: A look back and into the future". *ACM Computing Surveys,* ISSN: 03600300. https://doi.org/10.1145/2333112.2333115.

Xie, Zhongbin, e Shuai Ma. 2019. "Dual-view variational autoencoders for semi-supervised text matching". Em *IJCAI International Joint Conference on Artificial Intelligence.* ISBN: 9780999241141. https://doi.org/10.24963/ijcai.2019/737.

Zavitsanos, Elias, Georgios Paliouras, George A. Vouros e Sergios Petridis. 2007. “Discovering subsumption hierarchies of ontology concepts from text corpora”. Em *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007.* ISBN: 0769530265. https://doi.org/10.1109/WI.2007.47.

Zhang, Jingtai, Jin Liu e Xiaofeng Wang. 2016. “Simultaneous Entities and Relationship Extraction from Unstructured Text”. *International Journal of Database Theory and Application,* ISSN: 20054270. https://doi.org/10.14257/ijdta.2016.9.6.15.

Zhang, Peng, e Wanhua Su. 2012. “Statistical inference on recall, precision and average precision under random selection”. Em *Proceedings - 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2012.* ISBN: 9781467300223. https://doi.org/10.1109/FSKD.2012.6234049.

Zhong, Bei, Jin Liu, Yuanda Du, Yunlu Liaozheng e Jiachen Pu. 2016. “Extracting Attributes of Named Entity from Unstructured Text with Deep Belief Network”. *International Journal of Database Theory and Application,* ISSN: 20054270. https://doi.org/10.14257/ijdta.2016.9.5.19.

Zouaq, Amal. 2011. “An overview of shallow and deep natural language processing for ontology learning”. Em *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances.* ISBN: 9781609606251. https://doi.org/10.4018/978-1-60960-625-1.ch002.

Parte III

ANEXOS

A

DETALHES DE RESULTADOS

A.1 ANEXO A - EXTRAÇÃO DOS TERMOS MAIS COMUNS

Os termos apresentados na primeira coluna da tabela seguinte são os que ocorrem de forma mais regular nos fólios analisados, encontrando-se o respetivo número de ocorrências na segunda coluna.

Termos extraídos por este método	Contagem dos termos
co’	745
parte	653
terra	495
comprido	487
poente	484
largo	481
semeadura	471
norte	454
varas	417
sul	385
nacente	382
leuara	322
hua	283
alqueires	267
terras	231
leira	203
outra	200
caminho	197
quatro	193
...	...

Tabela 2: Termos extraídos pelo método de extração dos termos mais comuns.

A.2 ANEXO B1 - EXTRAÇÃO DOS TERMOS PELO YAKE

A primeira coluna contém os termos que a ferramenta YAKE extraiu do rosto do fólio 241, organizados consoante o peso do termo no fólio (apresentado na segunda coluna), desde o termo mais relevante até ao menos relevante.

Termos extraídos	Peso dos termos
nacente	0.056741315788051354
Pedro	0.054832566807838
Jorge Fernandes	0.05336235647586864
chão	0.05144063993929594
Pero martins	0.03738985818118633
Pero	0.03606484220011592
herdeiros	0.035132862940741576
Bastião Gonçalves	0.03310980368167354
Baltasar Gonçalves	0.03240957098392927
André Gonçalves	0.030007017574193735
Anes	0.02961905537454751
Gaspar Gonçalves	0.02826649256336562
Domingos	0.025190699451129252
alqueires	0.02266436761120528
semeadura	0.017458591386231088
levará	0.014129818526722555
Maria Gonçalves	0.013860386450987345
Martim Gonçalves	0.012628187210026358
Gonçalves	0.011478994313208655
Domingos Anes	0.006229275496827573

Tabela 3: Termos extraídos pelo YAKE do rosto do fólio 241.

A.3 ANEXO B2 - EXTRAÇÃO DOS TERMOS PELO YAKE (PARTE 2)

A primeira coluna contém os termos que a ferramenta YAKE extraiu do verso do fólio 89, organizados da mesma maneira que o anexo anterior.

Termos extraídos	Peso dos termos
terras	0.029719206712436257
uai pera	0.029711873797125723
souto	0.02914318986329283
Caua	0.02887068995142616
gado derrador	0.02854610664494087
cazas terrejras	0.03606484220011592
foẏ Adega	0.022952579280308527
caza	0.0221690152612376
tres cortes	0.019843033862141507
sul	0.019197276588006015
cazas	0.017735212208990082
vinha	0.015967392931684546
poente	0.015355655434858234
terra	0.012237320411003167
Nacente	0.011086558717799098
Meza	0.010287981465754242
Norte	0.008681190896183294
caminho	0.007261032270830888
vinha pequena	0.006564795283734648
Antonio Machado	0.005339411584847115

Tabela 4: Termos extraídos pelo YAKE do verso do fólio 89.

A.4 ANEXO C - COMPARAÇÃO DE TERMOS EXTRAÍDOS

A tabela seguinte contém os termos extraídos pelo método de obtenção dos termos mais comuns, e através da utilização da ferramenta YAKE, na primeira e segunda coluna, respetivamente.

Termos da lista obtida pelo primeiro método	Termos da lista obtida pelo segundo método
parte	norte
terra	sul
semeadura	poente
norte	semeadura
varas	largo
alqueires	campo
caminho	caminho
terras	terra
casal	varas
partes	casal
campo	terras
casas	herdade
propriedades	herdades
lugar	tombo
casa	juramento
prazo	propriedades
parede	casas
camara	moinho
monte	rega
dia	alqueires
	casa
	estrada
	vila
	parede
	monte

Tabela 5: Comparação entre a lista obtida pelo primeiro método e pelo segundo método.

A.5 ANEXO D - DEFINIÇÕES DE TERMOS EXTRAÍDOS

A tabela seguinte apresenta as definições dos termos extraídos, segundo o Vocabulário de Bluteau (Barros 2018) (Bluteau 1712-28).

Termo	Definição segundo o Vocabulário de Bluteau
semeadura	"A acção de semear ou o que se semeou. Sementis, is Fem. Plin. Columel."
largo	"Estendido em largura."
campo	"Pedaço de terra cultivada."
caminho	"O espaço pelo qual se vay de hvm lugar a outro."
terra	"Região. Certo espaço de terra.", "Chão. Campo."
varas	"A vara Portugueza contèm palmos Geometricos ..."
casal	"Huma casa, ou duas numa fazenda: ou casa no campo cõ terras de pão, ..., também se chama de povoação campestre de poucas casas."
herdade	"Quinta, campo, terra, que se tem herdado de seus pays."
juramento	"Affirmação, ou negação, que se faz chamando a Deos por testemunha, explicitamente, nomeando-o pello seu nome, ou implicitamente, jurando pellas creaturas de Deos, em quanto resplandece em ellas sua bondade, poder, & sabedoria."
propriedades	"Bens de raiz, com dominio, & poder absoluto para os vender, empenhar, & dispor delles."
moinho	"Engenho q serve de moer trio, cevada, &c. Consta de roda , rodizio, pennas, pouso, corredoura, aguilhão, fegurelha, lobeto, rela, vielas, veyo, quelha, ou calha, ..."
rega	“Aguar. Verter água com regador, ou outro instrumento.”
alqueires	"Medida de todo o genero de grãos."
casa	"Morada de casas, edificio, em que vive huma familia com seus moveis, & alfayas, amparada das injurias do tempo."
estrada	"Caminho público, por onde todos passaõ, a pé, a cavallo, em coche &c ..."
villa	"Povoação aberta, ou cercada, que nem chega a Cidade, nem he tão pequena, como Aldea."
parede	"Obra de pedra, & cal, que divide huma casa da outra."
monte	"Terra maninha, a monte, que está de monte, sem cultivo."

Tabela 6: Definições dos termos segundo o Vocabulário de Bluteau.

A.6 ANEXO E1 - RELAÇÕES ENTRE ELEMENTOS DA ONTOLOGIA

A tabela 7 contém as relações entre elementos da ontologia, que o algoritmo *Apriori* acompanhado de Regras de Associação considera mais relevantes, bem como o seu valor de *lift*.

Primeiro termo da relação	Segundo termo da relação	Valor de lift
casal	propriedades	4.828087167070217
caminho	casas	2.9540740740740743
campo	terras	2.739010989010989
campo	largo	2.3205320033250207
largo	varas	2.304099664631525
campo	terra	2.115980024968789
alqueires	semeadura	2.0770833333333334
casas	varas	2.0691251596424007
herdade	terra	1.9837312734082395
caminho	varas	1.9768103448275862
alqueires	terras	1.826007326007326
caminho	semeadura	1.7724444444444443
caminho	terra	1.75501872659176
semeadura	varas	1.7428400383141762
alqueires	caminho	1.7170555555555556
semeadura	terra	1.6907095297544734
caminho	largo	1.6658104738154615
estrada	semeadura	1.6443576388888888
alqueires	largo	1.6160847880299252
largo	semeadura	1.616084788029925
alqueires	terra	1.5766125676238034
caminho	terras	1.5338461538461539
casas	terras	1.5216727716727716
largo	terra	1.5178439668618715
estrada	terra	1.5169709737827715
alqueires	varas	1.5040948275862067

Tabela 7: Tabela de relações entre elementos segundo as Regras de Associação.

A.7 ANEXO E2 - RELAÇÕES ENTRE ELEMENTOS DA ONTOLOGIA (PARTE 2)

A seguinte tabela contém as relações a que o algoritmo *Apriori* acompanhado de Regras de Associação atribui um valor de *lift* superior a 2.0, ou seja, relações que o algoritmo considera muito relevantes. Na terceira coluna da tabela é também efetuada uma análise destas relações, no sentido de explicar a sua relevância.

Primeiro termo	Segundo termo	Análise da relação
casal	propriedades	um casal é um conjunto de casas e um tipo de propriedade
caminho	casas	caminhos tendem a ligar diferentes propriedades entre si, o que leva a que muitas casas confrontem com caminhos
campo	terras	um campo é um tipo comum de terra ou terreno
campo	largo	a largura de um campo que é um tipo de terreno é um atributo deste
largo	varas	a unidade de medida de um terreno é varas
campo	terra	um campo é um tipo de terra
alqueires	semeadura	alqueires são utilizados para medir cereais semeados
casas	varas	as casas mencionadas nos fólios eram medidas em varas

Tabela 8: Análise a relações com mais de 2.0 de lift.

A.8 ANEXO F1 - OBTENÇÃO DE INSTÂNCIAS DE PROPRIEDADES

Este anexo contém o *output* da primeira de três fases do processo de extração de instâncias de propriedades.

```
[
  'TM-F0001v.txt:TM-F0005.txt',
  [
    'partem', 'do', 'nacente', 'pella', 'estrada', 'que', 'uay', 'pera', '/',
        'a', 'igreia', 'de', 'são', 'Payo', 'e', 'das', 'mais', 'partes', '
       estão', 'dentro', 'enteiras', 'deste', 'Mesmo', 'casal.'
  ],
  [
    'as', 'casas', 'de', 'Soutulho', 'em', 'que', 'viuem', 'Maçio', 'Gomez',
       'e', 'Matheus', 'glz', 'terréas', 'e', 'Colmaças', '/', 'tem', "hua'",
        'casa', 'e', "hu'", 'curral', 'tudo', 'em', "hu'", 'corrume,'
  ],
  'Titulo da Camara de São Paỷo termo da Villa de Melgaço / em que ha as
     propriedades'
]
```

A.9 ANEXO F2 - OBTENÇÃO DE INSTÂNCIAS DE PROPRIEDADES (PARTE 2)

Este anexo contém o *output* da segunda de três fases do processo de extração de instâncias de propriedades.

```
[
  'TM-F0001v.txt:TM-F0005.txt',
  'Ho campo do terreo iunto as casas contra o Nacente cerquado sobre sỹ ao
      longo do caminho, Leuara de semeadura cinquo alqueires e meo de Centeo,
      tem ao Redor quatorse castanhejros /',
  {
    'Comprimento': ' oitenta e duas varas',
    'Largura': " cinquoenta e quatro he boa' terra tem da agoa sobredita."
  },
  [
    'de semeadura cinquo alqueires e meo de Centeo, tem ao Redor quatorse
        castanhejros /'
  ],
  {
    'Norte': "co' este Casal, tem de Comprido oitenta e duas uaras e de largo
         cinquoenta e quatro he boa' terra tem da agoa sobredita.",
    'Nascente': "co' este Casal, tem de Comprido oitenta e duas uaras e de
        largo cinquoenta e quatro he boa' terra tem da agoa sobredita.",
    'Sul': "co' este Casal, tem de Comprido oitenta e duas uaras e de largo
        cinquoenta e quatro he boa' terra tem da agoa sobredita.",
    'Poente': "com terras do mostejro de Paderne, e das mais partes co' este
        Casal, tem de "
  },
  "Ho campo do terreo iunto as casas contra o Nacente cerquado sobre sỹ ao
      longo do caminho, Leuara de semeadura cinquo alqueires e meo de Centeo,
      tem ao Redor quatorse castanhejros / parte do Norte com caminho e do
      Poente com terras do mostejro de Paderne, e das mais partes co' / este
      Casal, tem de Comprido oitenta e duas uaras e de largo cinquoenta e
      quatro he boa' terra tem / da agoa sobredita.",
  'campo do terreo',
  "Titulo da outra parte do Casal do Soutulho que foỹ emprazado / a Inez anes
      , pello Arcebispo Dom Frey Bartholameu dos Martires / aos dous dias do
      Mez de Julho de Mil quinhentos sesenta e sette annos \ufeff Pera que
      ella Ines Anes o possuisse en sua Vida e podesse Nomear Segunda pessoa
      Somente, pagasse / de foro vinte alquejres de Pão meado Milho e centeo e
       des almudes de Vinho e hua' Galinha pera a / aiuda dos quais vinte
      alqueires pagão os herdejros de Mario anes dous alquejres por resão do
      Campo / do prado, que possuem, que he pertença. doutra parte do Casal do
       Soutulho, que possuhio a dita Maria / anes de que atras se faz Mencão,
      He neste Prazo a Segunda Vida pella condicão / delle) derradejra
      Bartholameu Gomes Sangrado filho que foy da dita Ines anes, o qual casal
       / anda diuidido entre diuersas pessoas conuem a saber Antonio Lopez
```

```
        escriuão em Valladores, Miguel / Goncalues, Gil Gonçalues e Matheus gonç
        alues moradores no Mesmo Soutulho, e Fernão go’calues / e Gonçallo
        Esteues de Sante, o Padre Affonso da Lama Miguel affonso Marcos aluares
        de / Couello, os herdejros de Rui Montejro dos Barrejros conuem a saber
        Hyeronimo Roiz’ do Lagendo e / Costanca do Lagendo, e João Gomes e Gil
        gonçalues morador en Soutulho e João Gomez morador em São / Payo,
        Francisco Lourenço de Oriais e Rodrigo Seara de Sante Macio Gomes, e
        Maria Gomes filha / de Janebra Gomes, João Bieites de Sante e sua Maỷ
        Isabel alures’ Gonçallo Domingues genro de Affonso goncalues alfaiate
        Morador en sante, o qual casal tem as propriedades"
    ]
```

A.10 ANEXO F3 - OBTENÇÃO DE INSTÂNCIAS DE PROPRIEDADES (PARTE 3)

Este anexo contém o *output* da última de três fases do processo de extração de instâncias de propriedades.

```
[
  'TM-F0023v.txt:TM-F0027.txt',
  'Campo da Barroca',
  'Campo ',
  ' Nouenta e Noue varas',
  ' setenta e oỹto,',
  [
    'de semeadura oỹto alquejres de centeo pouco mais ou Menos, he terra
       fraqua tem da dita agoa'
  ],
  'e  com terras do Mostejro de Paderne,',
  "co' o Regueiro",
  'con terras que dizem deste casal ser dizimo a Deos, tem desanoue Carualhos
     pellos Comaros.',
  'com terras do Mostejro de Paderne,',
  "o Campo da Barroca que uay ter ao Porto do Atalho, que tem de Comprido
     Nouenta / e Noue uaras, e de largo setenta e oỹto, leuara de semeadura o
     ỹto alquejres de / centeo pouco mais ou Menos, he terra fraqua tem da
     dita agoa parte do sul co' o Regueiro / e do Norte e Nacente com terras
     do Mostejro de Paderne, e do poente con terras que dizem deste casal ser
      / dizimo a Deos, tem desanoue Carualhos pellos Comaros.",
  "Titulo do Cazal dos Barrocos, Sito em Penço termo de / Valadares, que foy
     de Gregorio do Barro, que o trouxe / per titulo de Prazo, que lhe fez
     Dom Frey Enrique Sendo Bispo de Ceita, pera elle e duas pessoas depoz
     elle a quatro de Nouembro de Mil / quinhentos e dous, pagasse de foro ao
      Snor' Arcebispo e sua Camara de São Payo per São Miguel / de Setembro
     de cada hu' anno dez alquejres de pão Meado milho e Centeo e sette
     almudes de / vinho; e duas galinhas Não constou quem era Vida no prazo,
     aprezentouo João Esteues \ufeff da Rocha morador no lugar do Barro da
     dita freiguesia de São tiago de Penço, possuem / com elle Gregorio
     Esteues, João Alexandre, Rodrigo Afonso esteuão Mouro, Maria esteuez, /
     Maria Glz', Pero esteuez, xpouão Vaz, Lourenço Esteuez e tem as
     ppriedades"
]
```

A.11 ANEXO G - OBTENÇÃO DE INSTÂNCIAS DE TÍTULOS

Este anexo contém o *output* do processo de extração de instâncias de títulos.

```
[
  "Titulo do Cazal do Gial, que trouxe Gildo Gial, Sito na / freyguezia de Sã
     o Martinho d'Aluarenga digo de Aluaredo, / que possuem Francisco Alres'
     e Affonso Vaz', e Ine alres', Mulher não Cazada, Ruy Gonçalues filho que
      ficou de João Artur e / os mais herdejros do dito João Artur não
     mostrarão prazo e dicerão que não sabião que nunqª o ouuesse, pagasse de
      foro deste casal ao snor Arçº e sua Camera de São Payo en cada hu' anno
      / quatro alqueires de Pão meado milho e Centeo, e hua' lamprea, e as
     propriedades são as",
  "Titulo do Cazal do Gial, que trouxe Gildo Gial, Sito na / freyguezia de Sã
     o Martinho d'Aluarenga digo de Aluaredo, / que ",
  "m Francisco Alres' e Affonso Vaz', e Ine alres', Mulher não Cazada, Ruy
     Gonçalues filho que ficou de João Artur e / os mais herdejros do dito Jo
     ão Artur não mostrarão prazo e dicerão que não sabião que nunqª o
     ouuesse, pagasse ",
  " deste casal ao snor Arçº e sua Camera de São Payo en cada hu' anno /
     quatro alqueires de Pão meado milho e Centeo, e hua' lamprea, "
]
```

A.12 ANEXO H - ELEMENTOS ONTOLÓGICOS

A tabela seguinte contém os elementos constituintes da ontologia obtida, bem como os seus tipos.

Nome do elemento	Tipo do elemento
Emprazador	Classe
Título	Classe
Propriedade	Classe
Vinha	Subclasse
Outros Terrenos	Subclasse
Vinha **é uma** Propriedade	Relação Taxonómica
Outro Terreno **é uma** Propriedade	Relação Taxonómica
Propriedade **faz parte de um** Título	Relação Não Taxonómica
Título **possui** Emprazadores	Relação Não Taxonómica
Nome	Atributo Simples
Preço	Atributo Simples
Tipo	Atributo Multivariado
Cava	Atributo Simples
Semeadura	Atributo Simples
confrontações	Atributo Composto
Fronteira Norte	Atributo Simples
Fronteira Nascente	Atributo Simples
Fronteira Poente	Atributo Simples
Fronteira Sul	Atributo Simples
Dimensões	Atributo Composto
Largura	Atributo Simples
Comprimento	Atributo Simples

Tabela 9: Elementos constituintes da ontologia.

NB: place here information about funding, FCT project, etc in which the work is framed. Leave empty otherwise.