PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE ESTRUTURAS ONTOLÓGICAS DE DOMÍNIO DA WIKIPÉDIA EM LÍNGUA PORTUGUESA

CLARISSA CASTELLÃ XAVIER

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof. Vera Lúcia Strube de Lima

Porto Alegre
2010

Dados Internacionais de Catalogação na Publicação (CIP)

X3e Xavier, Clarissa Castellã.
Extração de estruturas ontológicas de domínio da wikipédia em língua portuguesa / Clarissa Castellã Xavier. – Porto Alegre, 2010.
101 p.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientadora: Prof. Vera Lúcia Strube de Lima.

1. Informática. 2. Ontologia. 3. Wikipédia.
4. Processamento da Linguagem Natural. I. Lima, Vera Lúcia Strube de. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Extração de Estruturas Ontológicas de Domínio da Wikipédia em Língua Portuguesa", apresentada por Clarissa Castellã Xavier, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional , aprovada em 08/03/10 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – Orientadora — PPGCC/PUCRS

Profa. Dra. Renata Vieira – — PPGCC/PUCRS

Prof. Dr. Thiago Alexandre Salgueiro Pardo – — USP

Homologada em...../...../....., conforme Ata No. ..... pela Comissão Coordenadora.

Prof. Dr. Fernando Gehm Moraes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320 3611 Fax (51) 3320 3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Para meus amores:

Letícia, fonte infinita de alegria e inspiração.

Rodrigo, o melhor companheiro de jornada.

E AGORA?

Há críticos que, em vez de me julgarem pelo que sou,
julgam-me pelo que eu não sou.
É como quem olhasse um pessegueiro e dissesse:
"Mas isso não é um trator!"

Mário Quintana

AGRADECIMENTOS

Primeiramente agradeço a minha orientadora, professora Vera Lúcia Strube de Lima, pelos conselhos, ensinamentos e atenção recebida.

Agradeço a Letícia, minha filha nascida durante o mestrado, por toda alegria que trouxe em minha vida.

Ao Rodrigo, meu marido, por entender e apoiar minha opção pela vida acadêmica.

A Ana, minha mãe, pelo apoio e em especial por cuidar da Letícia para que eu pudesse dar andamento ao mestrado. É impossível esquecer as duas no sofá da sala, aguardando que eu retornasse da faculdade para amamentar.

Meus colegas do Grupo de Pesquisa em PLN da PUCRS, particularmente à Tatiane e Larissa, cujo apoio foi essencial para a conclusão das disciplinas do mestrado e a Mírian e o Douglas, que me proporcionaram muitos momentos alegres em 2009.

A Faculdade de Informática da PUCRS, excelência em ensino e pesquisa, a quem devo minha formação acadêmica.

Por fim, meu agradecimento ao CNPq pelo auxílio financeiro que possibilitou a realização deste trabalho.

EXTRAÇÃO DE ESTRUTURAS ONTOLÓGICAS DE DOMÍNIO DA WIKIPÉDIA EM LÍNGUA PORTUGUESA

RESUMO

A necessidade crescente por ontologias e a dificuldade em construí-las manualmente vêm gerando iniciativas em busca de métodos para a construção automática e semi-automática de ontologias. A Wikipédia, contendo uma grande quantidade de conteúdo organizado, livremente disponível e cobrindo uma extensa faixa de assuntos, mostra-se uma fonte interessante para extração de estruturas ontológicas. Neste trabalho propomos um método semi-automático para a extração de estruturas ontológicas de domínio a partir da estrutura de categorias da Wikipédia em português. Para validar o método proposto, realizamos um estudo de caso no qual foi implementado um protótipo gerando uma estrutura ontológica do domínio Turismo. Os resultados obtidos foram avaliados através da comparação da estrutura ontológica gerada com um mapeamento de referência, apresentando-se promissores, comparáveis aos encontrados na literatura para outros idiomas.

Palavras Chave: ontologias, Wikipédia, extração semi-automática de ontologias.

DOMAIN ONTOLOGICAL STRUCTURES EXTRACTION FROM WIKIPEDIA IN THE PORTUGUESE LANGUAGE

ABSTRACT

The increasing need for ontologies and the difficulty of its manual creation generates initiatives that propose methods for automatic and semi-automatic ontology construction. Wikipedia has demonstrated to be a very interesting source for ontologies extraction, due to the large amount of organized content in it, being freely available and covering a wide range of issues. In this work we propose a semi-automatic method of domain ontological structures extraction from Wikipedia's categories structure. To validate the method, we have conducted a case study in which we implemented a prototype generating a Tourism ontological structure. The results were evaluated by comparing them with a golden map of the generated ontological structure. The results are promising and comparable to those found in the literature for other languages.

Keywords: ontologies, Wikipedia, semi-automatic ontologies extraction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

HTML - *HyperText Markup Language*

OWL – *Ontology Web Language*

PLN – Processamento da Linguagem Natural

URI - *Uniform Resource Identifier*

URL - *Uniform Resource Locator*

XML - *EXtensible Markup Language*

W3C - *World Wide Web Consortium*

SUMÁRIO

1. INTRODUÇÃO

Embora ontologia seja um termo de criação moderna, remete a uma disciplina que foi inicialmente sistematizada e definida na filosofia, como “estudo do ser enquanto ser”, por Aristóteles, com o objetivo de fornecer sistemas de categorização para organizar a realidade. Outras áreas foram se apropriando do termo e da disciplina e, mais recentemente, a Inteligência Artificial passou a fazer uso do mesmo. Pesquisadores da Ciência da Computação adaptaram o termo aos seus próprios jargões, de acordo com Gruber em [GRU93] “*uma ontologia é uma especificação formal e explícita de uma conceitualização de um domínio*”.

O consórcio W3C[1] caracteriza ontologia como a definição dos termos utilizados na descrição e na representação de uma área do conhecimento. Sucintamente, o W3C coloca que as ontologias devem prover descrições para os seguintes tipos de conceitos:

- Classes (ou “coisas”) nos vários domínios de interesse.
- Relacionamentos entre essas “coisas”.
- Propriedades (ou atributos) que essas “coisas” devem possuir.

As ontologias auxiliam *software* e agentes humanos a se comunicar, provendo conhecimento sobre um domínio comum e compartilhado [MAE02]. Elas são utilizadas em diferentes aplicações, tais como resolução de problemas, tradução automática, expansão de consultas através de taxonomias, classificação de documentos baseados em aprendizado supervisionado e semi-supervisionado, entre outras [SUC08].

Construir manualmente ontologias, conforme [MAE02], é um processo oneroso, tedioso e propenso a erros. Além disso, o número de ontologias de domínio disponíveis é extremamente pequeno [HEP06], sendo em número menor ainda em língua portuguesa [LIM07].

Web Semântica[2] é o projeto capitaneado pelo W3C que pretende embutir inteligência e contexto nos códigos utilizados para confecção de páginas *Web*, de modo a melhorar a forma com que programas interagem com estas páginas e também possibilitar um uso mais intuitivo da *Web* por parte dos usuários [SOU04]. Ontologias são a base da *Web* Semântica.

---

1 http://www.w3.org/TR/2003/WD-webont-req-20030203/

2 Página oficial do projeto: http://www.w3.org/2001/sw/

A Wikipédia, de acordo com a definição dada pelo projeto que a mantém[3], é uma enciclopédia multilíngue, *online,* livre, sem fins lucrativos, colaborativa, criada em 15 de janeiro de 2001. Em dezembro de 2009 a Wikipédia contava com mais de 526.000 artigos em língua portuguesa e mais de 3.117.000 artigos em língua inglesa.

Os artigos da Wikipédia estão organizados em uma hierarquia de categorias e esta, com suas subcategorias, pode ser entendida como um grafo representando "*uma rede conceitual com relações semânticas não especificadas*"[4] [PON07a]. Este grafo, conforme [ZES07], compartilha diversas propriedades com outras redes semânticas lexicais, como, por exemplo, a WordNet[5]. Por este motivo, pode ser utilizado como um recurso em aplicações onde outras redes semânticas são tradicionalmente empregadas.

Diversos trabalhos [HEP06, VÖL06, WU07, AUE07a, PON07a, PON07b, SCH07, NAS08, PON08, SUC08, SY08, ZIR08] vem propondo a extração de estruturas ontológicas da Wikipédia. A enciclopédia *online* tem se mostrado uma fonte muito interessante para extração de estruturas ontológicas, visto que conta com milhões de entradas, centenas de milhares de colaboradores e milhões de artigos revisados [SPI08], cobrindo uma extensa faixa de assuntos, sendo uma das mais importantes coleções de conteúdo geradas por usuários [MIK08] disponível livremente.

Desta forma, o uso crescente, a dificuldade de criação manual e a carência de ontologias disponíveis, particularmente em língua portuguesa, bem como a riqueza de conteúdo disponibilizado pela Wikipédia, nos motivam a propor uma pesquisa voltada à extração de ontologias da Wikipédia. Dentro deste contexto, emerge a seguinte questão de pesquisa: "É possível extrair estruturas ontológicas de domínio em português a partir da estrutura de categorias da Wikipédia?"

Buscando responder a esta questão, realizamos o levantamento teórico, através do estudo das principais fontes sobre a construção de ontologias, propostas de extração de ontologias a partir da Wikipédia e, mas especificamente, construção de estruturas ontológicas a partir da estrutura de categorias da Wikipédia.

Krötzsch *et al.* [KRÖ05] apresentam a seguinte ideia ao introduzir o conceito de classe em um ontologia: "*Classes podem ser comparadas com categorias da Wikipédia: elas descrevem coleções de objetos e podem ser organizadas hierarquicamente, por exemplo ator é subclasse de pessoa. Como na Wikipédia, a herança múltipla e até mesmo ciclos são permitidos na hierarquia de classes*". Tal percepção, trouxe a estrutura

---

3 Definição disponível em http://pt.wikipedia.org/wiki/Wikipédia

4 Visto que a relação entre categoria e subcategoria não é necessariamente do tipo *is-a*.

5 http://WordNet.princeton.edu/

de categorias da Wikipédia como foco para o estudo.

Estudos preliminares, detalhados em [XAV09a] e [XAV09b], voltados à extração de uma estrutura taxonômica e relações de localização a partir da categoria Turismo da Wikipédia em português, nos levaram à proposta do método semi-automático para extração de estruturas ontológicas de domínio da Wikipédia em língua portuguesa apresentado nesta dissertação.

Para validar o método proposto desenvolvemos um estudo de caso e nesse contexto implementamos um protótipo produzindo uma estrutura ontológica do domínio Turismo contendo classes, instâncias e relações do tipo *is-a* e *located-in.* Os resultados obtidos foram avaliados através das seguintes métricas: Precisão, Abrangência e Medida-F. Para permitir esta avaliação, um modelo de referência (G*olden Mapping*) foi elaborado manualmente a partir da estrutura da categoria Turismo da Wikipédia em português, revisado e refinado por uma linguista.

Os resultados do estudo de caso foram promissores. As medidas extraídas na avaliação mostraram que a estrutura ontológica gerada pelo protótipo aproximou-se da estrutura de referência, demonstrando a viabilidade da extração de estruturas ontológicas de domínio em português a partir das categorias da Wikipédia, através do método proposto. Tais resultados são comparados aos encontrados na literatura, e servem de ponto de partida para considerações e trabalhos futuros.

O texto desta dissertação está organizado em 7 capítulos. O Capítulo 2 reúne a fundamentação sobre temas fundamentais de nosso estudo: ontologias, construção de ontologias e Wikipédia. O Capítulo 3 relata trabalhos correlatos ao presente estudo.

O núcleo desta dissertação encontra-se no Capítulo 4, que apresenta nossa proposta de método semi-automático para extração de estruturas ontológicas de domínio a partir das categorias da Wikipédia. O Capítulo 5 descreve um estudo de caso avaliando o método proposto no capítulo anterior. Os resultados alcançados no estudo de caso são apresentados e comentados no Capítulo 6. Concluímos o trabalho no Capítulo 7, tecendo as considerações finais, discutindo os resultados do estudo de caso, expondo as contribuições alcançadas e indicando trabalhos futuros.

O Apêndice A apresenta a lista das não-conformidades no mapeamento das relações *located-in* e *is-a* nas estruturas ontológicas de alinhamento e referência. O Anexo A apresenta um relatório referente à construção do modelo de referência utilizado na avaliação do estudo de caso. Este relatório foi elaborado pela linguista Susana de Azeredo e foi incluído para enriquecer o entendimento da construção do modelo de referência a partir da estrutura da categoria Turismo da Wikipédia.

2. FUNDAMENTOS

Neste capítulo trazemos a fundamentação teórica de nossa pesquisa, apresentando temas fundamentais de nosso trabalho: ontologias, construção de ontologias e Wikipédia.

2.1. Ontologias

Seres humanos, em geral, utilizam a linguagem natural para se comunicar e criar modelos a respeito do mundo. Entretanto, a linguagem natural, bastante ambígua, não é adequada para a construção de modelos na Ciência da Computação, dando lugar a linguagens formais quando o objetivo é especificar computacionalmente modelos do mundo [MAE02].

Mesmo assim, o compartilhamento e reuso da informação na Computação apresenta problemas. Quando diferentes sistemas utilizam termos distintos para descrever a mesma informação, surge uma lacuna semântica entre a representação sintática da informação e sua conceitualização [MAE02]. Ontologias surgem como uma alternativa para endereçar este problema.

Uma definição usualmente utilizada para ontologia é a de Gruber, apresentada em [GRU93]: "*uma ontologia é uma especificação formal e explícita de uma conceitualização de um domínio*". Tal definição enfatiza dois pontos chave [DAV06]:

- A conceitualização é formal e, de tal modo, permite o raciocínio por computadores.
- A ontologia deve ser projetada para um domínio particular de conhecimento.

A Figura 1 ilustra o sistema desenvolvido por Guarino [GUA98], que em contraste com a visão de Gruber, classifica as ontologias de acordo com o objeto de conceitualização, sugerindo que os diferentes tipos de ontologias sejam desenvolvidos de acordo com seu nível de generalidade.

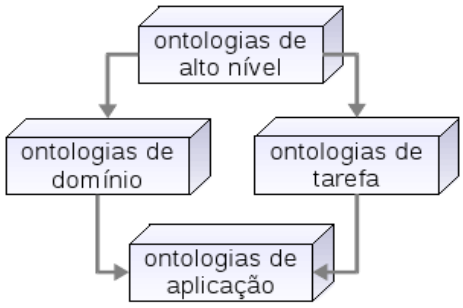


Figura 1 – Diferentes tipos de ontologias e seus relacionamentos - adaptação de [GUA98].

Caracterizamos os tipos de ontologia definidos por Guarino, baseados em

[MAE02]:

- Ontologias de alto nível: descrevem conceitos gerais como espaço, tempo, assuntos, objetos, eventos ou ações, entre outros, todos independentes de um problema ou domínio particular.
- Ontologias de domínio: descrevem o vocabulário relacionado a domínios genéricos através da especialização de termos introduzidos nas ontologias de alto nível.
- Ontologias de tarefa: descrevem o vocabulário relacionado a tarefas e atividades genéricas, especializando as ontologias de alto nível.
- Ontologias de aplicação: tipo mais específico. Seus conceitos geralmente correspondem a papéis desempenhados por entidades de domínio enquanto executam uma certa atividade.

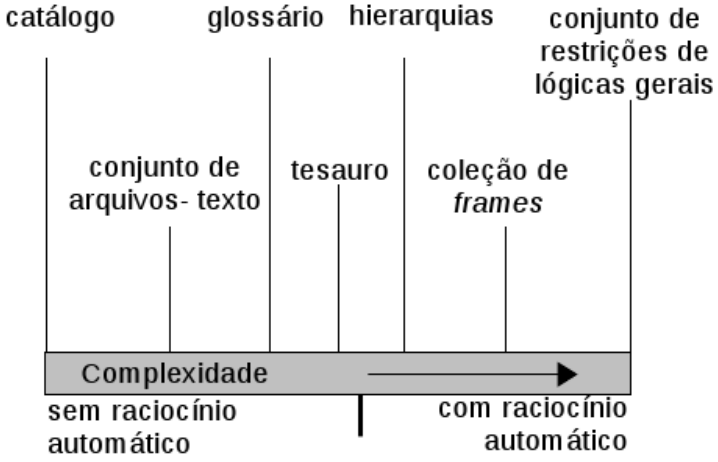


Figura 2 - Classificação das ontologias de acordo com sua complexidade - adaptação de [SMI01].

Smith e Welty classificam as ontologias de acordo com sua complexidade, conforme apresentado na Figura 2. Nesta classificação, varia o grau de formalismo e expressividade de cada representação. Todos os artefatos desta classificação objetivam estabelecer um vocabulário compartilhado que permita a troca de informações entre grupos de trabalho ou indivíduos [BRE05].

No presente trabalho utilizaremos os termos “ontologia” e “estrutura ontológica” intercambiadamente, e adotaremos, para ontologia, uma abordagem mais aberta, que pode remeter a uma terminologia dotada de relações semânticas simples.

Em termos práticos, ontologias descrevem conceitos (também denominados classes), relações (propriedades), instâncias (também conhecidas por indivíduos) e axiomas. Descrevemos a seguir estas primitivas de representação, baseados em [PER99]:

- Conceitos (classes): principais entidades de uma ontologia, representam um

conjunto de indivíduos do domínio. Em suma, um conceito pode ser qualquer coisa sobre algo que é dito e, portanto, pode ser a descrição de uma tarefa, função, ação, estratégia, raciocínio, etc.

- Instâncias (indivíduos): elementos particulares associados a conceitos do domínio.
- Relações: representam o tipo de interação entre conceitos de um domínio. Podem expressar, por exemplo, hierarquia (do tipo especialização) como *is-a*, agregação ou composição como *part-of* , localização como *located-in*.
- Axiomas: sentenças lógicas que definem a semântica dos conceitos e das suas relações.

Noy e McGuiness, em [NOY01], afirmam que "*desenvolver uma ontologia inclui definir as classes da ontologia; arranjar as classes em uma taxonomia (hierarquia de superclasses e subclasses); definir propriedades e descrever valores permitidos para essas propriedades; e preencher os valores para as propriedades das instâncias*". Os autores destacam três regras fundamentais no projeto e construção de uma ontologia [NOY01]:

- Regra 1: Não existe maneira correta de se modelar um domínio. Existem sempre várias alternativas e a melhor solução quase sempre depende da aplicação que se tem em mente e os acréscimos que são possíveis de serem previstos.
- Regra 2: O desenvolvimento de uma ontologia é necessariamente um processo iterativo. Depois de elaborada uma versão inicial de uma ontologia, ela deve evoluir, sendo certa a necessidade de revisão da ontologia inicial. Esse processo iterativo deve continuar durante todo o ciclo de vida da ontologia.
- Regra 3: É necessário sempre lembrar que uma ontologia é um modelo da realidade do mundo e os conceitos da ontologia devem refletir essa realidade, de tal forma que os conceitos da ontologia devem ser próximos aos objetos (físicos ou lógicos) e aos relacionamentos do domínio de interesse.

2.1.1. Construção de Ontologias

A construção de ontologias, também conhecida como aprendizagem de ontologias ou *ontology learning,* remete à integração de diferentes disciplinas [MAE01]. O desafio da construção de ontologias é preencher a lacuna entre o mundo dos símbolos da linguagem natural e o mundo dos conceitos, que essencialmente são abstrações do pensamento humano [CIM06]. A construção de ontologias é um problema que vem sendo abordado por diversos autores e trabalhos, sendo tipicamente referenciada como o gargalo da

aquisição do conhecimento (*knowledge acquisition bottleneck)* [CIM06].

Segundo Buitelaar *et al.* [BUI05] diversos aspectos diferenciam a construção de ontologias de outras atividades relacionadas à aquisição de conhecimento:

- A construção de ontologias é uma atividade multidisciplinar, particularmente em razão à sua forte conexão com a W*eb* Semântica, atraindo pesquisadores de áreas como a lógica, representação de conhecimento, filosofia, banco de dados, aprendizado de máquina e PLN, por exemplo.
- No contexto da W*eb* Semântica, a construção de ontologias está primariamente concentrada na aquisição de conhecimento "de" e "para" conteúdo na *Web,* tendo por interesse a massa de dados heterogênea da internet.
- Os métodos de avaliação desta área vêm sendo rapidamente adaptados para o uso da construção de ontologias.

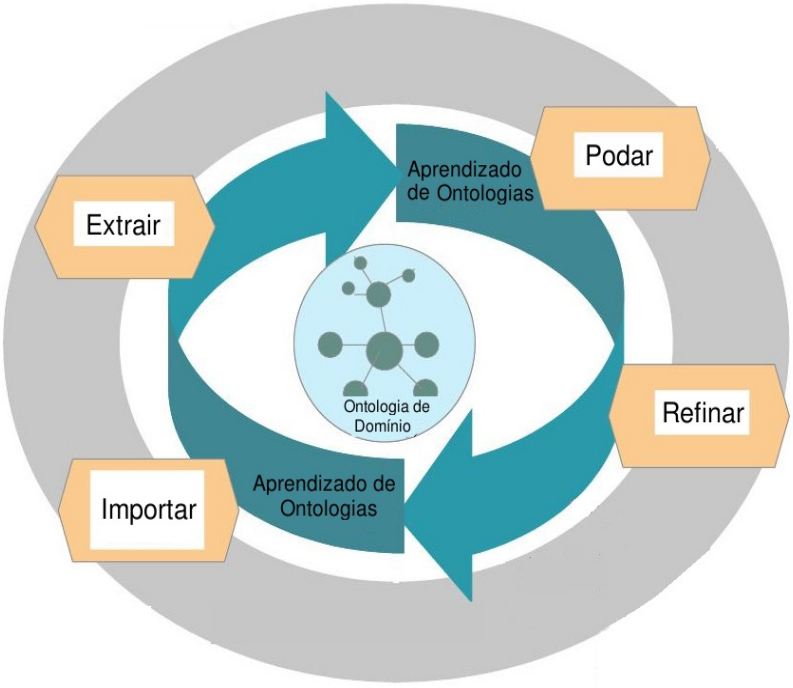


Figura 3 - Passos do processo de construção de ontologias - adaptação de [MAE01].

O ciclo de construção de ontologias proposto por Maedche [MAE01], ilustrado na Figura 3, tem por objetivo facilitar a construção e manutenção de ontologias pelo engenheiro de ontologias [MAE01]. Os passos do processo retratado neste ciclo são [MAE01]:

- Importação (e reuso): nesta etapa ontologias existentes são importadas e reutilizadas através da fusão de estruturas pré-existentes ou através de regras de mapeamento definidas entre estruturas existentes e a ontologia que será construída.
- Extração: nesta fase as partes principais da ontologia que está sendo construída

são modeladas.

- Poda: o esboço da nova ontologia é podada, a fim de ajustar a nova ontologia tendo em vista sua finalidade primordial.
- Refinamento: toma por base a ontologia de domínio elaborada nas etapas anteriores, e a conclui adicionando granularidade.

A extração consiste na fase de modelagem e aprendizagem da ontologia [MAE01]. A extração automática é vital para o sucesso da engenharia de ontologias, visto que lida com o gargalo da aquisição de conhecimento, ou seja, com a dificuldade em capturar e construir bases de conhecimento.

Os métodos automáticos realizam a extração de ontologias computacionalmente, sem interferência humana, enquanto que métodos semi-automáticos necessitam alguma forma de interferência do engenheiro de ontologias durante sua execução. Lau *et al*. em [LAU09] argumentam que, mesmo que a construção automática de ontologias de domínio ideais seja um campo em aberto, métodos de extração automática podem auxiliar engenheiros de ontologia a construir ontologias mais rapidamente e com maior qualidade.

Neste trabalho propomos um método que atua nesta etapa do processo de construção de ontologias.

2.2. Wikipédia

Em 1995, Ward Cunning propôs uma linguagem de marcação (WikiML) com o objetivo de simplificar a tarefa de criação e modificação de páginas *Web*, em particular de *links*, visto que o HTML se mostrava excessivamente complexo para o uso por usuários não-técnicos acostumados com o formato texto padrão das mensagens da época [BUF08]. Wikis são meios estabelecidos para a criação colaborativa, versionamento e publicação de artigos que buscam simplificar o processo de criação e manutenção de conteúdo de uma comunidade de leitores e, ao mesmo tempo autores [AUE07a].

Uma Wiki (forma simplificada para “WikiWikiWeb”, derivada da expressão havaiana “wiki wiki” que significa rápido, ligeiro) é essencialmente uma coleção de sítios *Web* conectados via *hyperlinks* [SCH07]. Um sistema Wiki é um sistema simples de gerenciamento de conteúdo, feito especialmente para permitir que o leitor possa alterar e melhorar livremente este conteúdo [VÖL06].

A Wikipédia, exemplo mais conhecido de Wiki público [BUF08], é uma enciclopédia *on-line*, criada por Jimbo Wales e Larry Sanger em janeiro de 2001 e desenvolvida por uma comunidade de usuários. Seu conteúdo cresce exponencialmente com a adição constante de artigos por seus colaboradores em todo planeta [SYE08]. Este crescimento

é atribuído a um misto de tecnologias e um processo de participação aberto [SPI08].

Porto Alegre

Origem: Wikipédia, a enciclopédia livre.
(Redirecionado de **Porto alegre**)

Porto Alegre é a capital do estado do Rio Grande do Sul. Pertence à mesorregião metropolitana de Porto Alegre e à microrregião de Porto Alegre. É localizada junto ao Guaíba, no extremo sul do Brasil, a 2.027 quilômetros de Brasília.

A cidade constituiu-se a partir da chegada de casais açorianos portugueses na primeira metade do século XVIII. No século XIX contou com o influxo de muitos imigrantes alemães e italianos (também recebeu imigrantes árabes e poloneses).

O feriado de Porto Alegre é o dia 2 de fevereiro, dia de Nossa Senhora dos Navegantes, festa religiosa mais popular da cidade. É uma das capitais estaduais no Brasil onde o índice de desenvolvimento humano é o mais elevado [9].

É a maior região metropolitana do sul do país, e a quarta do Brasil, com 3.959.807 habitantes (IBGE/2007).[10][2] Na capital gaúcha residem atualmente (2008) 1,43 milhão de pessoas, sendo a décima cidade mais populosa do Brasil de acordo com dados do IBGE.[2][11]

Em 2004, Porto Alegre foi eleita pela consultoria inglesa Jones Lang LaSalle uma das 24 cidades com maior potencial para atrair investimentos no mundo, e a única representante brasileira.[12].

Em 2001, a cidade foi sede da primeira edição Fórum Social Mundial, evento agora itinerante, que enfoca as questões sociais do mundo atual sob a perspectiva da esquerda política. Foi sede deste evento também em 2002, 2003 e 2005.

Índice [esconder]

1 História
2 Geografia
2.1 Bairros
2.2 Clima
2.3 Meio ambiente
2.3.1 Problemas ambientais
2.4 Parques e outras áreas públicas
2.5 Cidades vizinhas
2.5.1 Distâncias rodoviárias
3 Política
3.1 Câmara municipal
3.2 Prefeitos
4 Economia
4.1 Comércio
5 Demografia
5.1 Criminalidade
6 Cultura
6.1 Turismo
6.2 Esportes

Município de Porto Alegre

"A capital dos Gaúchos"
"Mui Leal e Valerosa"

Brasão Bandeira

Hino

Aniversário	26 de março
Fundação	1772
Gentílico	*porto-alegrense*
Lema	"Leal e Valerosa Cidade de Porto Alegre"
Prefeito(a)	José Fogaça (PMDB)

Localização

Figura 4 – Fragmento de artigo[6].

O conteúdo da Wikipédia está disponível na internet como páginas HTML estáticas, onde cada página possui um botão de edição que pode ser utilizado para atualizar seu conteúdo. O sistema que administra a Wikipédia mantém um histórico de edição de cada página e alerta usuários administradores quando páginas inscritas em uma lista de artigos vigiados sofrem alterações [SPI08].

Artigos[7] são páginas que contêm informações sobre algum assunto. O corpo do artigo pode conter informações suplementares como tabelas, imagens, mensagens em outras línguas, mensagens para os outros contribuintes da Wikipédia, referências a outros artigos da enciclopédia e *hyperlinks* para sítios externos. Artigos mais longos geralmente são divididos em seções ou subseções e possuem um índice [SCH07]. A Figura 4 apresenta um recorte de um artigo.

Cada artigo possui um registro contendo identificação e data/hora de criação e edições, número de referências à outras entradas da enciclopédia, número de revisões,

6 http://pt.wikipedia.org/wiki/Porto_alegre

7 http://pt. wikipedia.org/wiki/Artigo

número de entradas que se referem ao artigo, referências à outros artigos por mês e a indicação do período em que o artigo foi marcado como esboço[8] [SPI08].

Dados gerais	
País:	Alemanha
Estado:	Baden-Württemberg
Região Administrativa:	Karlsruhe
Distrito:	distrito urbano
Coordenadas geográficas:	49° 24' 44" N 08° 42' 36" E
Altitude:	116 metros acima do nível do mar
Área:	108,83 km²
População:	142.993 (31 Dez. 2005)
Densidade populacional:	1.314 hab./km²
Código postal:	69001–69126
Código telefônico:	06621
Endereço da prefeitura:	Marktplatz 10 69117 Heidelberg
Website:	sítio oficial
Prefeito:	Dr. Eckart Würzner

Figura 5 - Exemplo de *infobox.*[9]

Infobox é um modelo padrão da Wikipédia, que contém uma tabela com informações básicas sobre a entidade descrita no artigo onde esse *infobox* está inserido. Por exemplo, *infoboxes* de artigos que descrevem países costumam conter informações como o nome do país na língua nativa, sua capital e área. A Figura 5 apresenta o recorte do *infobox* do artigo referente à cidade de Heidelberg na Wikipédia em língua portuguesa.

Os documentos da Wikipédia estão organizados em uma hierarquia de categorias, construída colaborativamente, com o objetivo de indexar os artigos [VÖL06]. A rede de categorias pode ser considerada uma *folksonomia,* ou seja, um sistema de marcação colaborativo que permite que os usuários categorizem as entradas da enciclopédia [STR06].

A organização dos artigos em categorias, segundo a própria Wikipédia[10], permite que eles sejam agrupados e que estes grupos também sejam categorizados. Quando um artigo pertence a uma categoria, ele contém um *link* para a página que descreve a categoria. Do mesmo modo, quando uma subcategoria pertence a uma categoria pai, ela irá conter um *link* para a página da categoria pai.

---

8 Esboço é um artigo que ainda não está completo (apenas iniciado) ou possui muito pouca informação.

9 http://pt.wikipedia.org/wiki/Heidelberg

10 http://en.wikipedia.org/wiki/Wikipedia:FAQ/Categorization

Figura 6 – Fragmento de página da categoria Hotelaria[11].

Cada página de categoria contém uma introdução que pode ser editada, e uma lista de *links* gerada automaticamente para artigos e subcategorias que pertencem à categoria. A Figura 6 apresenta um recorte de uma página de categoria.

O sistema de categorias não é uma hierarquia rígida ou uma árvore de categorias, já que cada artigo pode aparecer em mais de uma categoria, e cada categoria pode aparecer em mais de uma categoria pai. Tal organização em uma hierarquia de categorias, que não é uma estrutura estritamente arbórea, mas uma representação mais rica, permite que vários sistemas de categorização coexistam simultaneamente, na forma de um grafo que representa uma rede conceitual com relações semânticas não especificadas [PON07a].

O grafo de categorias da Wikipédia compartilha diversas propriedades com outras redes semânticas lexicais, como, por exemplo, a WordNet [MIL95, FEL98]. Por este motivo, pode ser utilizado como um recurso em aplicações onde outras redes semânticas são tradicionalmente empregadas [ZES07].

No tocante à confiabilidade de suas informações, [KIT08] relata que seu conteúdo

[11] http://pt.wikipedia.org/wiki/Categoria:Hotelaria

possui qualidade comparável a enciclopédias tradicionais, e que o vandalismo e imprecisões em seus artigos são freqüentemente revertidos em questão de minutos. Em 14 de dezembro de 2005 a revista britânica Nature realizou uma pesquisa comparativa, referente a 50 artigos científicos, entre a Wikipédia e a Enciclopédia Britânica[12]. Quarenta e dois artigos da Wikipédia foram analisados por especialistas e o resultado da comparação obteve números semelhantes entre as duas:

- Inconsistências por verbete (média):
  - Wikipédia **= 4**; Enciclopédia Britânica **= 3**;
- Erros graves:
  - Wikipédia **= 4**; Enciclopédia Britânica **= 4**;
- Incorreções factuais, omissões e afirmações falsas
  - Wikipédia **= 162**; Enciclopédia Britânica **= 123**;

Hepp, Bachlechner e Siorpaer publicaram uma análise quantitativa da Wikipédia, em [HEP06], provando que as *URIs* das entradas da Wikipédia são confiáveis como conceitos de uma ontologia e mostrando que entradas disponíveis na Wikipédia podem ser utilizadas como elementos de uma ontologia. O estudo verificou que apenas 3% da amostra tornaram-se páginas de desambiguação durante o período avaliado. Além disso, 94 de 100 entradas permaneceram estáveis, podendo ser utilizadas como fontes de dados sem maiores problemas.

Wu e Weld, em [WU07], mencionam os seguintes atributos que tornam a Wikipédia ideal para extração de dados:

- Todos os conceitos importantes possuem um identificador único, a URI (*Uniform Resource Identifier*) do artigo.
- *Infoboxes* podem ser utilizados como fonte de dados para treino, permitindo aprendizado supervisionado.
- Listas e categorias fornecem um sistema de tipos simples e uma taxonomia rudimentar, respectivamente.
- Páginas de redirecionamento podem ser utilizadas para identificação de sinônimos.
- Páginas de desambiguação podem ser utilizadas na geração de listas de candidatos para resolução de homônimos.
- Seu tamanho, grande o suficiente para fornecer o um bom número de informações.

No capítulo a seguir apresentamos trabalhos propondo a extração de estruturas ontológicas a partir da Wikipédia e, em particular, da sua estrutura de categorias.

[12] Disponível em http://www.nature.com/nature/journal/v438/n7070/full/438900a.html

3. TRABALHOS RELACIONADOS

Neste capítulo apresentamos trabalhos que relatam métodos para extração de estruturas ontológicas da Wikipédia, em especial aqueles que utilizam como fonte de dados as categorias da enciclopédia.

Iniciaremos apresentando um artigo que relata um experimento de extração de ontologias a partir dos modelos *infoboxes* e das categorias, através da aplicação de heurísticas. Em seguida, apresentaremos trabalhos de Strube, Ponzetto *et al.* cuja fonte primeira de dados é a estrutura de categorias da enciclopédia.

3.1. Extração a partir de *Infoboxes* e Categorias

YAGO: A Large Ontology from Wikipedia and WordNet [SUC08], de Suhaneck, Kasneci e Weikum, apresenta a ontologia YAGO (*Yet Another Great Ontology*), derivada da Wikipédia e da WordNet. A ontologia relatada possui aproximadamente 1,7 milhões de entidades[13] e 15 milhões de fatos[14], incluindo uma hierarquia *is-a*, assim como outras relações semânticas entre entidades. A Tabela 1 apresenta o número de entidades em YAGO.

Tabela 1 - Número de entidades da ontologia YAGO – adaptação de [SUC08].

Relações	92
Classes	224.391
Indivíduos	1.531.588
Pessoas	546.308
Locais	230.988
Instituições/companhias	57.893
Filmes	33.234

Os fatos que populam a ontologia foram extraídos do sistema de categorias e dos *Infoboxes* da Wikipédia, sendo combinados com as relações taxonômicas da WordNet. Segundo os autores do artigo, o método utilizado apresenta uma Precisão de 95% devido à heurística de extração utilizada através de checagem de tipos (*type checking*).

Os autores argumentam que *infoboxes* são processados muito mais facilmente que textos em linguagem natural, conforme demonstrado em [AUE07b] *apud* [SUC08]. As páginas de categorias (como, por exemplo, categoria Cantores Americanos) fornecem candidatos para entidades (por exemplo, Elvis Presley), candidatos para conceitos *is-a*

[13] Os autores denominam entidades todos os elementos da ontologia: relações, classes e indivíduos.

[14] Os autores denominam fatos as relações entre os elementos da ontologia.

*(*Elvis, CantorAmericano*)* e candidatos para relacionamentos, como, nacionalidade*(*Elvis, americano*)*.

A construção da ontologia YAGO é feita em dois estágios: primeiramente diferentes heurísticas são aplicadas à Wikipédia para extrair fatos e entidades candidatas, e é estabelecida a conexão entre Wikipédia e WordNet. O segundo passo é a aplicação de técnicas de controle de qualidade.

Visto que todos os fatos candidatos passam por um controle de qualidade, o sistema é bastante generoso na aplicação de heurísticas. Cada título de uma página da Wikipédia é um candidato a se tornar um indivíduo na ontologia. O algoritmo processa o XML da Wikipédia e aplica quatro heurísticas diferentes aos artigos:

1. Heurística de *Infobox*: Cada linha de um *infobox* contém um atributo e um valor. Os autores identificaram 170 atributos mais frequentes, para os quais foi designada manualmente uma relação, denominada Relação Alvo. Em princípio, cada linha de um *infobox* gera um fato. Por exemplo, os atributos "nascido" e "aniversário" geram a relação "Data de Nascimento".
2. Heurística de Tipos: Os autores classificam as categorias da Wikipédia como conceituais, administrativas, relacionais e temáticas. Para o sistema, apenas as categorias conceituais podem se tornar classes para os indivíduos da ontologia. Para distinguir as categorias conceituais das temáticas é feito um processamento linguístico dos nomes das categorias. Por exemplo, a categoria “Cidadãos naturalizados dos Estados Unidos” é quebrada em: pré-modificados (naturalizados), cabeça (cidadãos) e pós-modificados (dos Estados Unidos). Desta análise do sintagma nominal conclui-se que, se a cabeça do nome da categoria é uma palavra no plural, a categoria tende a ser conceitual. Após, utiliza-se apenas as categorias folhas da árvore de hierarquias da Wikipédia, e a hierarquia das classes é estabelecida pela WordNet. Em seguida, é feita a conexão entre a Wikipédia e a WordNet, ligando as classes superiores extraídas da WordNet às classes inferiores extraídas da Wikipédia.
3. Heurística de Palavras: Para estabelecer a relação “significa” (*means*), o sistema, além de explorar a WordNet, explora as páginas de redirecionamento da Wikipédia, buscando nomes alternativos para entidades através da busca por páginas de redirecionamento, o que resulta em um relacionamento do tipo Significa, como, por exemplo: "Einstein, Albert" significa “Albert Einstein”.
4. Heurística das Categorias: Categorias do tipo relacionamento fornecem informações sobre a entidade artigo. Por exemplo, a categoria Rios da Alemanha

informa que a entidade está localizada na Alemanha. Este tipo de informação é muito útil, principalmente em artigos que não possuem *infobox*. Assim, a heurística de categorias consiste em uma expressão regular, como "Rios|Montanhas em (.*)" e uma relação alvo, como *located-in.* Se o nome de uma categoria se enquadra na expressão regular, um fato novo é adicionado à ontologia.

3.1.1. Controle de Qualidade e Resultados

Um controle “canônico” (*canonicalization*) faz com que cada fato e cada referência a uma entidade sejam únicos, resultando que cada entidade é sempre referenciada pelo mesmo identificador em todos os fatos da ontologia.

A checagem de tipo pode ser utilizada de duas maneiras: redutiva, eliminando fatos que são implausíveis, e indutiva, adicionando fatos suplementares para que a ontologia se torne mais consistente. A checagem de tipo elimina todos os indivíduos que não possuem uma classe, assim como elimina os fatos que não respeitam as restrições de domínio e alcance de uma relação. Como resultado, um argumento de um fato em YAGO é sempre uma instância da classe requerida pelo relacionamento.

Tabela 2 - Precisão das heurísticas de YAGO – adaptação de [SUC08].

Heurística	**Precisão**
hasExpenses	100.0%
hasInflation	100.0%
hasLaborForce	97.67441%
during	97.48950%
ConceptualCategory	96.94342%
participatedIn	96.94342%
plays	96.94342%
establishedInYear	96.84294%
createdOn	96.84294%
originatesFrom	96.84294%
...	
WordNetLinker	95.11911%
...	
InfoboxType	95.08927%
hasSuccessor	94.86150%
...	
hasGDPPPP	91.22189%
hasGini	91.00750%
discovered	90.98286%

Este modelo baseado em heurísticas foi avaliado por juízes humanos que analisaram 5200 fatos, apresentando o seguinte resultado: 74 heurísticas tiveram

Precisão maior que 95%. A Tabela 2 listas as heurísticas que apresentaram melhor Precisão.

Atentamos para o fato que a heurística *located-in* não se encontra listada nesta tabela, o que nos leva a crer que a expressão regular implementada não apresentou resultados tão significativos como os demais. Além disso, os autores não informam o número total de heurísticas implementadas, de tal forma que não é possível precisar o peso de 74 heurísticas dentro do conjunto total.

3.2. Extração a partir das categorias da Wikipédia

Strube e Ponzetto conceituam a estrutura de categorias da Wikipédia como sendo uma rede semântica de conceitos organizada em um grafo acíclico dirigido, diretamente direcionado. Isto porque o software de edição da enciclopédia permite que os usuários criem ciclos, que, no entanto, de acordo com as orientações do guia de criação de categorias, devem ser evitados [PON07a]. A Figura 7, extraída de [PON07a], ilustra uma simplificação deste grafo, onde o direcionamento é abstraído.

Inicialmente as categorias da enciclopédia foram utilizadas pelos autores em [STR06] e [PON07a] como apoio a experimentos sobre computação da similaridade semântica[15] entre termos. Os resultados apresentados em [PON07a] mostraram que os resultados obtidos com o uso das categorias da Wikipédia no cálculo das relações semânticas correspondem mais aos resultados obtidos por juízes humanos do que o *baseline*, baseado no Google *counts*. Também mostraram que as categorias podem ser utilizadas em tarefas de resolução de correferência.

Partindo para um enfoque onde a estrutura de categorias passa a ser a principal fonte de dados, o artigo *Deriving a large scale taxonomy from wikipedia* [PON07b] relata um experimento onde é feita a extração de uma grande taxonomia contendo relações do tipo *is-a* e *not-is-a* a partir da estrutura de categorias da Wikipédia. A estrutura obtida contém mais de 105.418 ligações *is-a*. Os autores avaliaram o resultado através da comparação da taxonomia com a base de dados ResearchCyc[16] e via similaridade semântica com a WordNet, concluindo que os resultados obtidos são competitivos com ambas as ontologias.

---

[15] A similariedade semântica, neste contexto, indica o quanto dois conceitos são semanticamente distantes em uma rede ou taxonomia [PON07a].

[16] http://research.cyc.com/

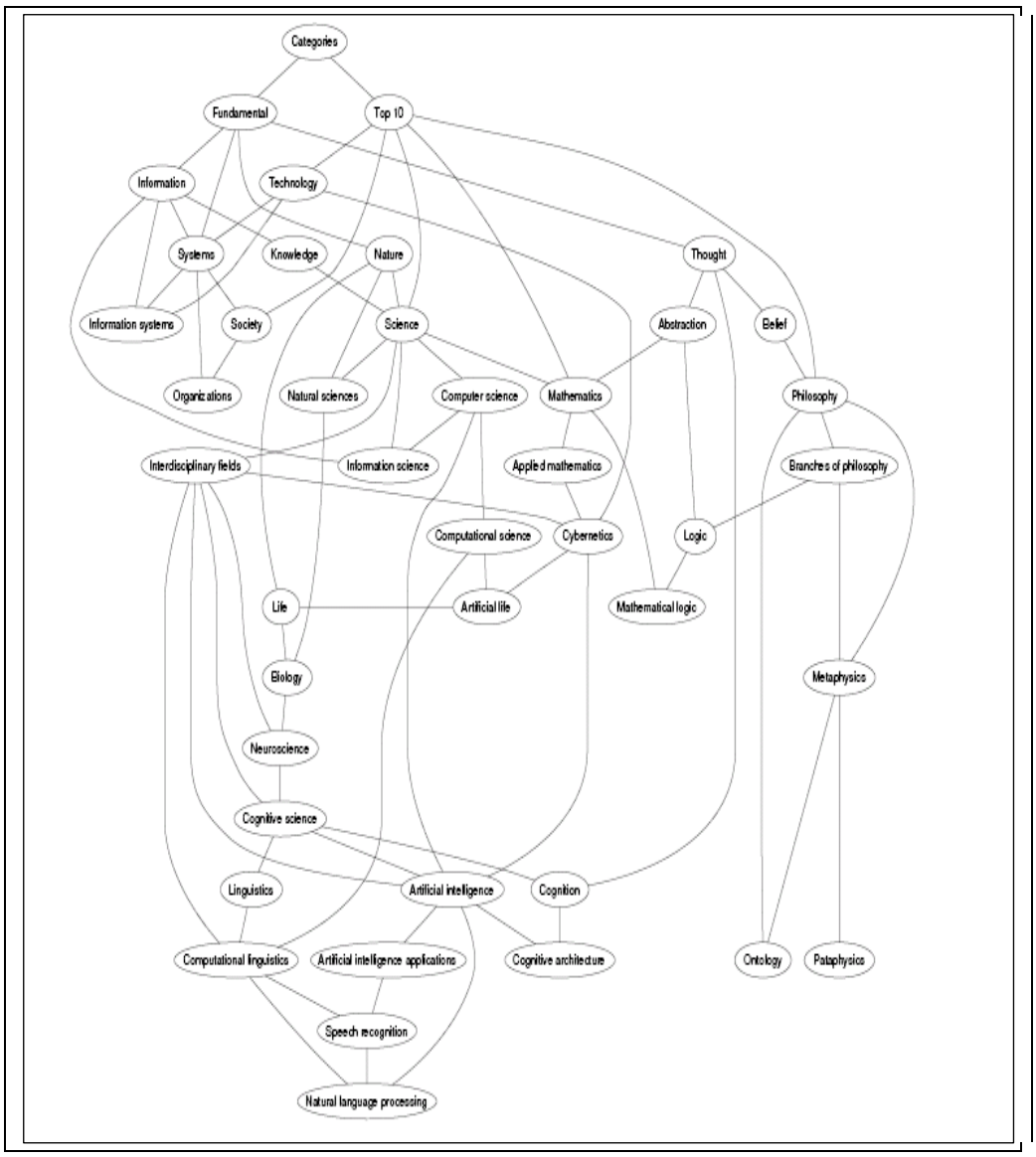


Figura 7 – Simplificação do grafo de categorias da Wikipédia – extraído de [PON07a].

Artigos posteriores, escritos por Ponzetto e Strube [PON08], Zirn, Nastase e Strube [ZIR08] e Nastase e Strube [NAS08] descrevem métodos para distinguir automaticamente classes e instâncias a partir da taxonomia relatada em [PON07b].

A seguir apresentaremos as abordagens utilizadas pelos autores nas tarefas de extração taxonômica e distinção entre classes e instâncias.

3.2.1. Extração da Taxonomia

Para extrair uma grande taxonomia da estrutura de categorias da Wikipédia foi realizado um experimento identificando relações de subsunção entre categorias, ou seja, relações *is-a* e *not-is-a.* Esta identificação foi realizada através de métodos baseados na conectividade da rede de categorias e comparação de padrões léxico-sintáticos, de acordo com os seguintes passos [PON07b]:

a) Eliminação das meta-categorias administrativas, por exemplo, categorias

sob *Wikipedia Administration* e categorias contendo em seu título as palavras *wikipedia, wikiprojects, template, user, portal, categories, articles* e *pages.*

b) Marcação como *is-refined-by* das relações entre as categorias cujos nomes correspondam ao padrão como o do exemplo: C1 = *Miles Davis Albums* e C2 =*Albums by Artist*.

c) Uso de métodos baseados na sintaxe:

- Atribuição do marcador *is-a* à ligação entre duas categorias que compartilham os mesmos lemas da cabeça do sintagma nominal (lemas lexicais principais). Por exemplo: *British Computer Scientists* e *Computer Scientists*.
- Atribuição do marcador *not-is-a* à ligação de uma categoria contendo o lema principal de outra categoria em uma posição diferente da principal. Por exemplo: C*rime comics* e C*rime*.

d) Uso de métodos baseados na informação estrutural da rede de categorias:

- Atribuição de marcador *is-a,* conforme o seguinte exemplo: a categoria *Microsoft* tem uma página homônima categorizada em E*mpresas listadas na Nasdaq* que tem o lema principal *Empresas*. *Microsoft* tem uma super categoria *Empresas de Computador e Videogame* com o mesmo lema principal. A ligação entre a *Microsoft* e *Empresas de Computador e Videogame* é rotulada como *is-a*.
- Atribuição do rótulo *is-a* para a ligação entre duas categorias se uma página é redundantemente categorizada abaixo das duas categorias. Por exemplo: *Ethyl Carbamate* é subcategoria de *Amide* e também de *Organig Compound(s)*. A regra infere que, por transitividade, *Amide is-a organic Compounds.*

e) Uso de padrões lexico-sintáticos em um corpus.

- Aplicação de padrões léxico-sintáticos em um grande corpus (Wikipédia em inglês e o Tipster corpus [PON07b] *apud* [HAR93]). Uso dos padrões definidos em [HEA92, CAR99] para identificar relações *is-a* e padrões definidos em [BER99, GIR06] para identificar relações de meronímia para identificar relações *not-is-a.* Para melhorar a abrangência destes padrões, quando o lema cabeça é identificado pelo Reconhecedor de Entidades Nomeadas descrito em [FIN05] como pertencendo à uma entidade nomeada, é utilizado todo o nome da categoria (por exemplo, *Brands* em *YUM! BRANDS*), senão apenas o cabeça (por exemplo, *Albuns* em *MILES DAVIS ALBUMS)*. Para assegurar a precisão na aplicação dos padrões, tanto a Wikipédia quanto o Tipster foram

preprocessados, para identificar sintagmas nominais, por um *pipeline* constituído do *POS tagger* estatístico baseado em trigramas [BRA00] e por um *SVM-based chunker* [KUD00].

Os padrões são utilizados para evidenciar as relações semânticas empregando uma estratégia de máximo de votos. Um par de categorias é rotulado positivamente com *is-a* no caso do número de casamentos *(matches)* de padrões positivos ser maior que o número de casamentos de negativos. Adicionalmente, os padrões são utilizados para filtrar as relações *is-a* criadas pelos métodos baseados em conectividade (itens f e g).

f) Propagação das relações encontradas pelo método anterior, propagando as relações *is-a* para todos que sigam o seguinte padrão: visto que *Microsoft is-a companies Listed in NASDAQ*, infere-se que *Microsoft is-a Multinational Companies (*superclasse com mesmo lema principal).

g) Atribuição do rótulo *is-a* para as ligações baseadas na transitividade - todas as categorias em uma cadeia *is-a* são conectadas entre si por ligações *is-a*, como por exemplo, *Fruit is-a Crops* e *Crops is-a Edible Plant*, inferindo a relação *Fruits is-a Edible Plant*.

A taxonomia obtida ao final do experimento constou de 105.418 ligações do tipo *is-a* e foi avaliada através da sua comparação com a ontologia ResearchCyc[17]. Foram avaliados 85% dos pares gerados contendo conceitos correspondentes na base referência. Esses pares foram avaliados através de uma busca na ResearchCyc quando o conceito denotado pela subcategoria da Wikipédia era uma instância-de *(#$isa*) ou generalizado-por *(#$genls)* pelo conceito denotado em sua superclasse. O resultado desta busca foi tomado como a classe semântica do par de categorias (*is-a* ou *not-is-a*) e utilizado na avaliação. Desta forma foram computadas as medidas de Precisão, Abrangência e Medida-F.

A Tabela 3 apresenta os resultados da aplicação de toda a sequencia de métodos propostos.

Tabela 3 - Extrato dos resultados de [PON07b].

	R	P	F1
Todos métodos	89,1	86,6	87,9

3.2.2. Distinção entre classes e instâncias

Em *Distinguishing between Instances and Classes in the Wikipedia Taxonomy* [ZIR08]

[17] research.cyc.com

são propostos os seguintes métodos para efetuar a distinção entre classes e instâncias:

a) Baseados na estrutura
   - Atribuição do rótulo Classe para cada categoria que possuir pelo menos duas hiponímias e para cada categoria que possuir exatamente uma hiponímia, se ela for mais de uma hiponímia ela própria.

b) Baseados no nome da categoria
   - Com o uso de um Reconhecedor de Entidades Nomeadas e um classificador baseado em *Condition Randon Fields* [FIN05], etiqueta-se o título da categoria com rótulos *pessoa, localização, organização* e *outro*. Se a maioria das palavras do título for etiquetada como *outro* ela é rotulada classe, senão é rotulada como instância.
   - Segundo as regras de atribuição de nomes da Wikipédia, palavras que constituem parte de uma entidade nomeada são capitalizadas iniciam em maiúscula. O método é preprocessar a primeira palavra do título com o classificador e, se não for reconhecida como Entidade Nomeada, mudar todas suas letras para minúscula; filtrar as palavras funcionais (preposições, artigos, etc); e analisar as palavras restantes no título. Após este tratamento, as palavras que iniciarem em maiúscula são instâncias.
   - Títulos de categorias que representam instâncias devem estar no singular. Se uma das palavras principais (determinadas pelo *Standford Parser*) está marcada como nome no plural, a categoria é definida como classe, senão como instância.
   - Segundo as normas da Wikipédia, artigos devem ser colocados em categorias com o mesmo nome. Por este motivo, categorias que contém artigos com o mesmo nome são marcados como instâncias.

Atentamos para o fato que o método baseado na estrutura não foi claramente explicado pelos autores.

Para avaliar os resultados obtidos, o repositório ResearchCyc foi utilizado como *Gold Standard*. Esta estrutura contém uma marcação distinguindo entre *#$Individual* (indivíduo) e *#$SetOrCollection* (conjunto ou coleção) para cada entrada. Foram encontrados 7860 conceitos em comum entre o conjunto avaliado e a referência (44,45% indivíduos e 55,65% conceitos), sendo este o conjunto avaliado. A combinação dos métodos utilizados para distinguir entre classes e instâncias, com base na precisão individual de cada um deles, resultou em um algoritmo que identifica as instâncias com Precisão de 90,92% e 84,52% de Abrangência [ZIR08].

No artigo *WikiTaxonomy: A Large Scale Knowledge Resource* [PON08], os autores criam a taxonomia, conforme os passos de [PON07b], e a partir daí seguem critérios similares aos descritos em [ZIR08] para realizar a distinção entre classes e instâncias, de acordo com o algoritmo a seguir:

- Sendo L uma categoria
  - Se nenhuma página é intitulada L e a lema principal do título de L é plural, L é Classe.
  - Senão, L é capitalizada e foi reconhecida pelo Reconhecedor de Entidades Nomeadas como Entidade Nomeada, L é Instância.
  - Senão, se não existe página entitulada L, L é Classe.
  - Senão, se o *head* de L é plural, L é Classe.
  - Senão, se L não possui sub-categoria, L é Classe.
  - Senão, se L é capitalizado , L é Instância.
  - Senão aplica-se o padrão: L é Classe

[PON08] relata os seguintes resultados: foram classificadas 111.652 classes e 15.472 instâncias, com acurácia de 84,5% na comparação contra a base ResearchCyc.

A abordagem relatada em *Decoding Wikipedia Categories for Knowledge Acquisition* [NAS08] é decodificar automaticamente o título das categorias da Wikipédia e determinar as relações, classes e atributos neles embutidos, explorando os nomes de categorias e a estrutura de categorias como fontes de relações entre conceitos, induzindo as seguintes informações: instâncias de relações, tipos de relações e atributos de classes.

Os autores identificaram os seguintes tipos de informação nos títulos das categorias [NAS08]:

- Categorias contendo relações explícitas: indicam diretamente uma relação como *membro-de* (Membros do Parlamento Europeu), *causado_por* (Acidentes de avião causados por erros dos pilotos).
- Categorias contendo relações parcialmente explícitas: contém preposições que indicam relações semânticas, como em "*Villages in Brandesburg*" e "*conflicts in* 2000". Tal situação de ambiguidade, ou seja, os diferentes significados da preposição *in,* pode ser resolvida utilizando informações de entidades nomeadas, ou o grafo das categorias da Wikipédia: supercategorias de *Brandesburg* (geografia) e 2000 (século, anos) indicam os tipos de relação que o título da categoria contém.
- Categorias contendo relações implícitas: categorias cujo título são nomes complexos, mas não explicitam indicadores do tipo de relação. Exemplo: "*mixed*

martial arts television programs" possui 2 sintagmas: "*mixed martial arts*" e "*television programs*".

- Categorias classe-atributo: seguem o padrão "*x by y*", indicando generalização e atributos de classe. Depois de decodificar a informação do título de uma categoria, essa informação pode se propagar na rede de categorias. Categorias como "*albums by artist*" geralmente possuem subcategorias mais detalhadas, por exemplo *Miles Davis albums,* e são ligadas a páginas correspondendo a álbuns específicos.

As fases do processo de extração das relações semânticas e atributos de classe relatadas em [NAS08] são as seguintes:

a) Identificar o componente dominante: por exemplo, a categoria "C*hairmen for he County Councils of Norway*" possui três constituintes: *chairmen, county councils e Norway*, sendo o componente dominante *chairmen.*

b) Extrair relações: são selecionadas as páginas categorizadas em uma determinada categoria (tomadas como instâncias) e extraídas as relações delas com a categoria e relações presentes nos títulos das categorias, conforme descrito acima.

c) Extrair atributos de classe e instância: por exemplo, *Miles Davis Albums* é subcategoria de *Albums by Artist*. Neste caso, o algoritmo identifica a relação *Miles Davis is-a Artist,* identificando *Miles Davis* como instância e *Artist* como classe.

O processamento das categorias iniciou com a limpeza das categorias administrativas, resultando em uma rede de 197.667 categorias. Estas categorias foram processadas com o *POS tagger*, *parser* e Reconhecedor de Entidades Nomeadas desenvolvido pelo *Standord NLP group*[18], obtendo os seguintes números:

- Categorias contendo relações explícitas: 3.450
- Categorias contendo relações parcialmente explícitas e implícitas: 98.855
- Categorias classe-atributo: 7.564, sendo 840 classes com uma média de 2,27 atributos.

Tabela 4 - Resultados de [NAS08].

Tipo da Categoria	#categorias	#relações extraídas		Avaliação	
			P	manual ∩	manual U
relações explícitas	3,450	86,649			
caused-by, based-in, writen_by, ...	*2,152*	*43,938*	*-*	*94,37%*	*96,38%*
member_of	*1,398*	*42,711*	*24%*	*95,56%*	*97,17%*
relações parcialmente explícitas e implícitas	98,855	9,751,748			
is-a		*3,400,243*	*44,57%*	*76,45%*	*84%*
spatial		*3,201,125*	*39,69%*	*87,09%*	*97,98%*

A Tabela 4 apresenta os resultados da avaliação da comparação com a base

[18] http://www-nlp.stanford.edu/software/

ResearchCyc e anotação manual de relações. No caso dos falsos positivos da comparação com a ResearchCyc, foram selecionados aleatoriamente conceitos para anotação manual. Cada sub-conjunto de relações foi anotado independentemente por dois juízes, com dois escores: um para interseção (instâncias que ambos anotadores validam como corretas) e união (instâncias anotadas como corretas por apenas um anotador).

3.3. Considerações a respeito dos trabalhos apresentados

A pesquisa por trabalhos que apresentam mecanismos para a utilização da Wikipédia, assim como outras coleções de dados, na extração de estruturas ontológicas é um trabalho muito vasto, que permite uma série de discussões. Optamos, neste capítulo, por apresentar apenas as referências que apoiaram diretamente a construção do método proposto nesta dissertação.

O uso das heurísticas na extração de informação da Wikipédia, mostra-se interessante e facilmente transportável para a versão em língua portuguesa da enciclopédia. A combinação dos dados extraídos dos *infoboxes* com dados da WordNet para criação de ontologias, mostra-se promissora na língua inglesa. Entretanto, atualmente, este recurso ainda não aparenta ser útil para extração de dados da Wikipédia em português por dois motivos: *infoboxes* são pouco utilizados nesta versão da enciclopédia e as versões disponíveis da WordNet em língua portuguesa ainda não possuem uma quantidade significativa de conteúdo.

A estrutura de categorias da Wikipédia nos parece uma excelente fonte de dados para extração de uma estrutura ontológica, visto que a sua organização já sinaliza o relacionamento entre conceitos. Nos trabalhos de Strube e Ponzetto [STR06], [PON07a], [PON07b] as categorias são utilizadas para descrever conceitos. Nos trabalhos de Strube com Zirn e Nastase [ZIR08] e Nastase [NAS08] são relatadas análises mais detalhadas dos títulos, efetuando a extração de classes e instâncias. Verificamos que, também em língua portuguesa, muitas categorias abarcam em seu título relações e instâncias.

[PON07b] descreve com bastante clareza os passos seguidos para a identificação das relações *is-a* e *not-is* na estrutura de categorias. Entretanto, não deixa claro como foi realizado o uso de padrões léxico-sintáticos em um corpus nesta tarefa.

Também [ZIR08] não deixa claro como são utilizados os métodos baseados na estrutura, na distinção entre classes e instâncias, visto que o conceito utilizado de “classe sendo uma hiponímia ela própria” não é claro, nem ilustrado por exemplos.

A avaliação dos resultados encontrados nos trabalhos analisados foi realizada

através da comparação com outras estruturas e banco de dados. Infelizmente, ainda não dispomos em língua portuguesa de uma grande base de dados para validação de resultados, como a WordNet, fazendo com que procuremos soluções alternativas para avaliação do trabalho proposto nesta dissertação.

Apesar de apresentar excelentes números, a tabela de resultados das heurísticas de YAGO (Tabela 2) não lista os resultados da aplicação da heurística *located-in.* Esta ausência nos leva a crer que a expressão regular utilizada na heurística das categorias não apresentou resultados interessantes.

Não encontramos, nos trabalhos apreciados, nenhuma proposta relacionada à extração de ontologias de domínio. Todos tomaram por base o conteúdo integral da enciclopédia e não uma categoria específica. Além disso, os trabalhos analisados propõem a extração de conteúdo da enciclopédia em língua inglesa. A realização de um trabalho que extraia uma estrutura ontológica da versão em língua portuguesa deve levar em conta as características próprias desta versão, tais como o menor uso de modelos e as diferentes grafias de um mesmo termo, face às diferenças inerentes ao português brasileiro e europeu.

4. MÉTODO PARA EXTRAÇÃO DE ESTRUTURAS ONTOLÓGICAS DE DOMÍNIO DAS CATEGORIAS DA WIKIPÉDIA

Este capítulo constitui o núcleo desta dissertação, apresentando nossa proposta de método semi-automático para extração de estruturas ontológicas de domínio a partir das categorias da Wikipédia.

Estudos e experimentos relatados em [XAV09a] e [XAV09b], efetuando a extração de uma estrutura taxonômica e relações de localização a partir da categoria Turismo da Wikipédia em português, nos levaram ao método apresentado neste trabalho.

O método proposto encontra-se inserido na fase de extração do ciclo de aprendizado de ontologias [MAE01], apresentado na seção 2.1 deste trabalho. Nesta fase efetua-se a modelagem e aprendizagem de ontologias.

O método possui quatro etapas bem definidas, representadas na Figura 8:

- Extração da taxonomia;
- Identificação de relações, classes e instâncias;
- Formatação e unificação linguística;
- Geração da descrição OWL.

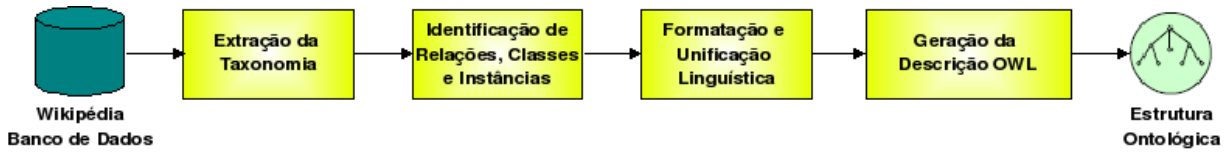


Figura 8 - Visão geral do método proposto.

A entrada do método é o banco de dados da Wikipédia, em particular as tabelas contendo o cadastro das categorias. Após a aplicação das quatro etapas detalhadas a seguir, obtém-se como saída uma estrutura ontológica de domínio, descrita em OWL, contendo classes, instâncias e relações.

O método baseia-se na identificação dos componentes da estrutura ontológica na nos títulos das categorias da Wikipédia. Primeiro é realizada a extração de uma taxonomia refletindo o cadastro da estrutura de categorias. Em seguida é feita uma análise de cada título desta estrutura taxonômica e destes títulos são extraídos novos relacionamentos, novas classes e instâncias gerando uma estrutura taxonômica

A seguir, detalharemos cada etapa.

4.1. Etapa 1: Extração da Taxonomia

O objetivo desta etapa é obter uma estrutura taxonômica, ou seja, uma hierarquia de conceitos, onde a relação hierárquica é estabelecida pela maneira como a estrutura de categorias foi organizada no banco de dados da Wikipédia e os conceitos são os títulos das categorias.

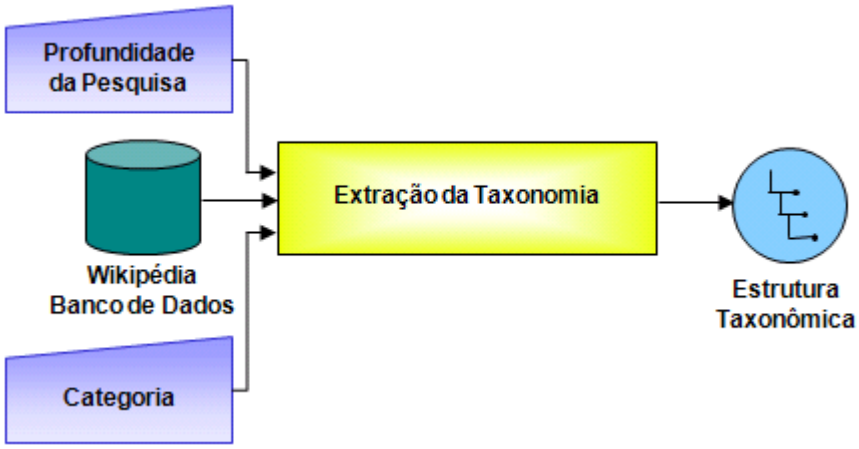


Figura 9 - Representação gráfica da primeira etapa.

A entrada de dados desta etapa, representada na Figura 9, é o banco de dados da Wikipédia, em particular as tabelas contendo os dados relacionados à estrutura de categorias. Também é preciso que o engenheiro de ontologias defina previamente qual categoria será pesquisada e a profundidade da pesquisa. A saída da etapa é uma seleção hierarquizada dos títulos das subcategorias, representando uma taxonomia.

O corpus da Wikipédia abrange diferentes campos do conhecimento e a organização do seu grafo de categorias viabiliza a ligação de conceitos que pertencem a domínios distintos. Visto que desejamos extrair uma estrutura de domínio, é preciso limitar a seleção das subcategorias a um determinado nível de profundidade, visando obter o maior número possível de conceitos, sem extrapolar o domínio, ou seja, garantindo que eles pertençam ao domínio que está sendo descrito.

→ Turismo por País
→ Turismo no Brasil
→Meio ambiente do Brasil
→ Energia no Brasil
→ Empresas de energia do Brasil
→ Empresas do setor elétrico do Brasil

Figura 10 - Recorte da estrutura de subcategorias de Turismo por País.

Para isso é preciso realizar previamente a análise da estrutura da categoria da qual serão selecionadas as subcategorias que formarão a taxonomia, e delimitar a profundidade da consulta. Por exemplo, a hierarquia das subcategorias de Turismo por País, ilustrada na Figura 10, nos leva para outro domínio do conhecimento quando aprofundado em mais de 3 níveis o escopo da seleção.

Neste caso, se optássemos por um nível menor de profundidade (por exemplo, apenas 2 níveis de profundidade) obteríamos uma estrutura taxonômica com um número muito limitado de conceitos, não utilizando plenamente o conteúdo disponível na Wikipédia.

A partir da definição prévia da profundidade da pesquisa, efetua-se a seleção, no banco de dados da Wikipédia, das subcategorias da categoria escolhida, em tantos níveis quantos forem adequados. Ressaltamos que o número de níveis a serem pesquisado na árvore de categorias para extração da taxonomia varia de acordo com o domínio, sendo informado manualmente de acordo com a característica da categoria a ser utilizada como base da extração.

4.2 Etapa 2: Identificação das Relações, Classes e Instâncias

Nesta etapa é realizada a análise dos conceitos presentes na taxonomia gerada na etapa anterior, e realizada a extração de relações, novas classes e instâncias. A Figura 11 ilustra estas atividades.

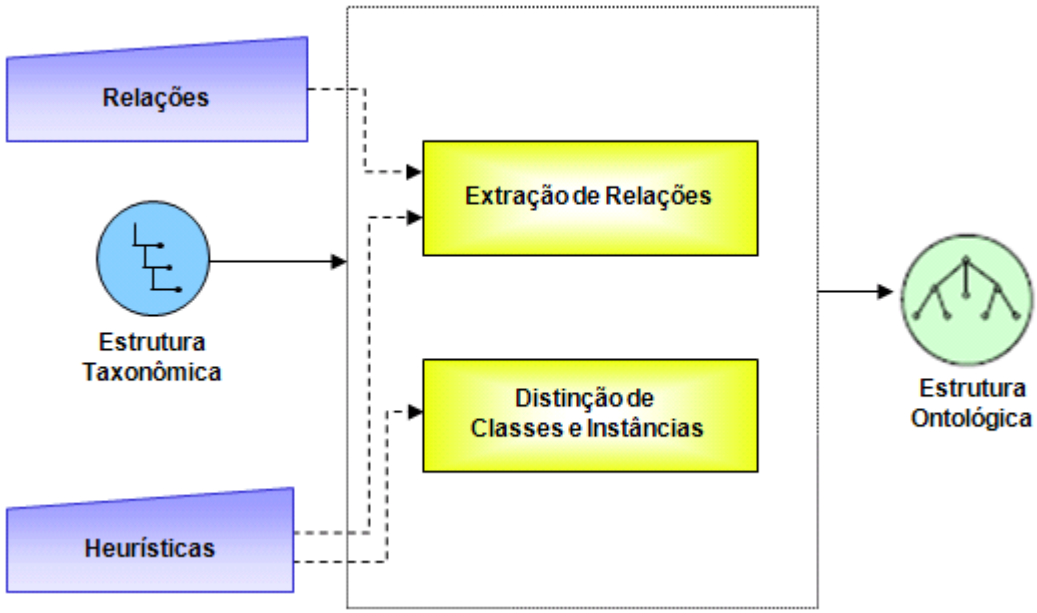


Figura 11 – Ilustração da segunda etapa.

Ao final desta etapa é gerada uma estrutura ontológica composta por classes, instâncias e relações. Para chegarmos a esta estrutura, é necessário definir previamente

as relações que serão extraídas, bem como as heurísticas a serem utilizadas para extrair as relações e distinguir classes e instâncias presentes nos títulos das classes da estrutura taxonômica gerada anteriormente.

Nesta fase, analisaremos cada conceito da taxonomia gerada na fase anterior, verificando se existe, embutida nele, uma relação semântica diferente da hiponímia. Neste caso, substituiremos a relação *is-a* pela nova relação e verificaremos se a classe permanece com o mesmo nome, ou se a partir da nova relação será criada uma nova classe e um conceito.

As atividades propostas nesta etapa foram inspiradas na Heurística das Categorias relatada em [SUC08], apresentada na seção 2.1 deste trabalho, e nos métodos descritos em [PON08], [ZIR08] e [NAS08] para distinguir automaticamente classes e instâncias a partir de uma taxonomia, apresentados na seção 2.2.

4.2.1. Extração das Relações

Segundo Gruber, "*relações são um conjunto de tuplas que representam um relacionamento entre objetos em um universo de discurso*" [GRU92].

Strube e Ponzetto afirmam que as categorias da Wikipédia não formam uma taxonomia com uma hierarquia de subsunções bem formada, mas um tesauro tematicamente organizado [PON07b]. O emprego exclusivo da relação de hiponímia (*is-a*) não descreve com fidelidade o relacionamento semântico entre os conceitos presentes na taxonomia extraída das categorias da Wikipédia. O uso de outras relações em conjunto com *is-a* é essencial para descrever com mais exatidão as ligações semânticas entre os conceitos. Por exemplo, a categoria Capitais da Ásia está cadastrada abaixo de Capitais (*is-a*), mas Filosofia está abaixo de Abstração e Crença (*deals-with*) e também abaixo de Humanidades (*is-a*) [PON07b].

A opção por quais relações devem ser extraídas está diretamente relacionada ao domínio que será representado pela estrutura ontológica. É necessário analisar a estrutura taxonômica obtida na etapa anterior e, a partir de seus conceitos, definir previamente quais relações representam os relacionamentos semânticos entre as classes.

Por exemplo, no domínio Turismo algumas categorias apresentam em seu título um relacionamento de localização, tal como "Jardins zoológicos da Alemanha", que pode ser melhor representado pela relação "Jardins zoológicos *located-in* Alemanha".

4.2.2. Distinção entre classes e instâncias

Instâncias representam os objetos do domínio sobre os quais reside nosso interesse, enquanto classes são interpretadas como conjuntos que contêm instâncias [HOR04]. Para caracterizar uma instância, utilizaremos os pontos apresentados por Miller e Hristea em [MIL06]: "*instâncias são nomes próprios, iniciando com maiúscula e instâncias são únicas, não podendo ser instanciadas elas próprias*".

Nesta etapa, para os conceitos em que verificamos, na etapa anterior, existir uma relação diferente da hiponímia, analisamos se é necessário criar novas classes e instâncias.

Por exemplo, identificamos na categoria "Jardins zoológicos da Alemanha" a relação "Jardins zoológicos *located-in* Alemanha". A partir daí, iremos caracterizar Jardins zoológicos como sendo uma nova classe e Alemanha como sendo uma instância da classe "local".

As heurísticas utilizadas para analisar os títulos das categorias devem ser definidas previamente. Isto porque são dependentes do domínio descrito pela estrutura ontológica que está sendo extraída, das relações que estão sendo identificadas e da análise dos títulos da estrutura de categorias utilizada como fonte.

4.3 Etapa 3: Formatação e Unificação Linguística

Figura 12 – Representação gráfica da terceira etapa.

Nesta etapa é realizada a formatação dos conceitos presentes na estrutura ontológica gerada na etapa anterior.

O primeiro passo é a unificação ortográfica dos conceitos, visto que a Wikipédia em língua portuguesa, segundo seu livro de estilo, não usa uma versão específica da língua comum, independentemente do seu país de origem. Deste modo, o mesmo termo pode estar cadastrado mais de uma vez com grafias distintas, com as diferenças inerentes ao português brasileiro e europeu.

Por este motivo realizamos a unificação da ortografia dos títulos de categorias no caso de a mesma palavra estar escrita de duas formas distintas como, por exemplo, “atracções” e “atrações". A ortografia brasileira foi definida como padrão, sendo usado, neste caso, “atrações”.

O segundo passo é a formatação dos títulos, para que eles possam ser representados em OWL ao final do método. Esta tarefa é realizada em três fases:

- Remoção dos caracteres especiais;
- Substituição de espaços por traço inferior (*underscore*);
- Conversão de todas as letras para minúsculas.

A Figura 13 apresenta um exemplo da execução desta etapa do método:

Atracções de Auckland → Atrações de Auckland → atracoes_de_auckland

Figura 13 – Exemplo de execução da terceira etapa.

4.4. Etapa 4: Geração da Descrição OWL

A meta desta etapa é gerar a descrição em OWL da estrutura ontológica obtida nos passos anteriores do método. Esta é a única fase do método independente do domínio descrito pelo estrutura ontológica gerada. A geração do arquivo OWL ao final do método permite que a estrutura ontológica extraída seja visualizada e refinada em editores de ontologias, bem como seja acessada automaticamente por outras aplicações.

A linguagem OWL é baseada em lógica de descrição, sendo desenvolvida com o propósito de facilitar a interpretação automática do conteúdo disponível na *Web* [MCG04], fornecendo vocabulário adicional em conjunto com uma semântica formal [PAN07]. OWL é recomendada pelo W3C desde janeiro de 2004 como o padrão do projeto *Web* Semântica.

Uma ontologia descrita em OWL consiste de classes, propriedades e instâncias [HOR04]. Conforme a indicação[19] do W3C, OWL destina-se a fornecer uma linguagem para descrever classes e as relações entre elas, que são inerentes aos documentos da *Web* e aplicações. OWL pode ser utilizada para:

- Formalizar um domínio pela definição de classes e propriedades destas classes.
- Definir instâncias e propriedades sobre elas.
- Raciocinar a respeito destas classes e instâncias até o grau permitido pela semântica.

[19] www.w3.org/TR/owl-guide/

4.5. Contribuições do Método

O objetivo do método é prover ao engenheiro de ontologias uma estrutura inicial para a extração de estruturas ontológicas de domínio a partir das categorias da Wikipédia. A base do seu funcionamento é a identificação dos componentes da estrutura ontológica nos títulos das categorias.

Embora baseado em propostas da literatura, o método reúne diferentes abordagens para construir o resultado. Por este motivo, trata-se de uma inovação, em especial em língua portuguesa, carente de ontologias a serem construídas para aplicações da *Web* Semântica.

O método é constituído por quatro etapas: na primeira etapa é realizada a seleção da estrutura taxonômica da categoria. Em seguida, na segunda etapa, são obtidas relações, novas classes e instâncias a partir da taxonomia gerada anteriormente. Na terceira etapa é realizada a normalização dos títulos das classes e instâncias e, finalmente, na quarta etapa, é gerado o arquivo OWL contendo a descrição da estrutura ontológica obtida.

No capítulo seguinte, apresentaremos um estudo de caso delineado para validar o método proposto.

5. ESTUDO DE CASO

Neste capítulo descrevemos um estudo de caso no qual foi gerada uma estrutura ontológica do domínio Turismo a partir da estrutura de categorias da Wikipédia em língua portuguesa, utilizando o método apresentado no capítulo anterior.

5.1. Descrição do Estudo de Caso

Com objetivo de validar o método proposto, foi desenhado um estudo de caso apoiado por um protótipo implementado em PHP[20], acessando o banco de dados MySql[21], gerando uma estrutura ontológica do domínio Turismo descrita em OWL.

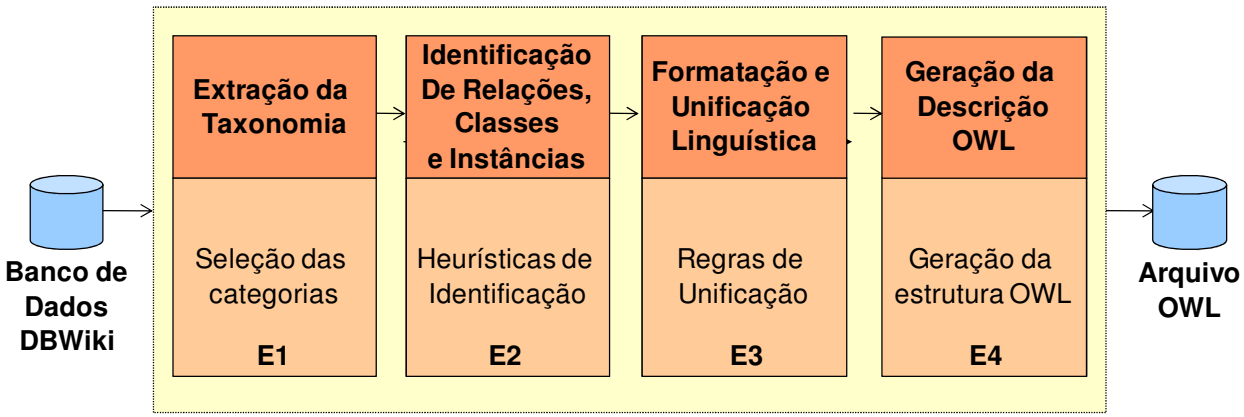


Figura 14 – Arquitetura do Protótipo.

A arquitetura do protótipo, ilustrada na Figura 14, seguiu as quatro etapas do método proposto. Na primeira etapa (E1) é realizada a seleção da estrutura taxonômica da categoria Turismo. Em seguida, na segunda etapa (E2), são obtidas relações, novas classes e instâncias a partir da taxonomia gerada na etapa anterior. Na terceira etapa (E3) é realizada a normalização do títulos das classes e instâncias e, finalmente, na quarta etapa (E4), é gerado o arquivo OWL contendo a descrição da estrutura ontológica obtida, que será utilizada na avaliação dos resultados.

O estudo de caso busca validar o método proposto através da extração e avaliação de uma estrutura ontológica de Turismo. O protótipo foi elaborado para trabalhar dentro deste domínio. Caso outro domínio de conhecimento fosse escolhido, seriam necessárias outras análise e escolhas nos pontos do processo em que há interferência para escolhas

[20] www.php.net

[21] Para a execução deste estudo de caso foi utilizada uma imagem do banco de dados da Wikipédia em língua portuguesa de 05 de janeiro de 2009, obtida no sítio de *downloads* do banco de dados da Wikipédia: http://download.wikimedia.org/backup-index.html

e modelagem da estrutura ontológica, cabendo uma nova prototipação.

Nas subseções a seguir, oferecemos a descrição mais detalhada das quatro etapas.

5.1.1. Etapa 1: Extração da Taxonomia

Nesta etapa realizamos a seleção dos conceitos de Turismo e sua organização em forma de taxonomia.

O corpus da Wikipédia abrange diferentes campos do conhecimento e a organização do seu grafo de categorias viabiliza a ligação de conceitos que pertencem a domínios distintos. Conforme o processo que estamos seguindo, realizamos uma análise do grafo das subcategorias de Turismo e, a partir desta observação, decidimos fixar a profundidade de nossa busca em três níveis, conforme ilustrado na Figura 15, buscando obter o maior número de conceitos possível, sem ultrapassar o domínio escolhido.

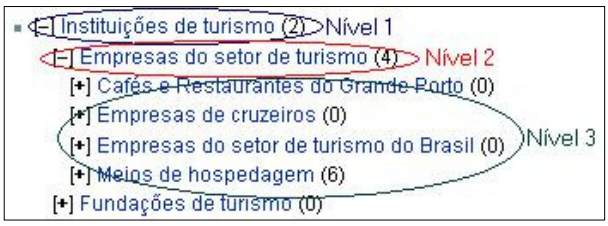


Figura 15 - Três níveis de subcategorização (categoria Turismo).

Tivemos por entrada as tabelas contendo informações a respeito das categorias e sua organização no banco de dados da Wikipédia[22]: *categorylinks, category e page*.

Selecionamos os títulos e ligações entre as subcategorias de Turismo, em três níveis, excluindo:

- Categorias do tipo “esboço”, ou seja, cujo conteúdo ainda não está completo (apenas iniciado) ou possui muito pouca informação. Estas categorias estão sinalizadas com o caractere "!" no início de seu título.
- Ligação[23] de uma categoria para ela mesma, evitando a presença de laços, buscando uma representação taxonômica mais correta.

[22] O esquema do banco de dados do MediaWiki encontra-se disponível em http://www.mediawiki.org/wiki/Manual:Database_layout para maiores informações.

[23] Estas ligações são adicionadas erroneamente pelos colaboradores da Wikipédia e são excluídas posteriormente na revisão de conteúdo.

Figura 16 - Recorte da estrutura taxonômica selecionada.

Ao final desta etapa, obtivemos uma taxonomia onde os conceitos são os títulos das categorias selecionadas e a relação hierárquica é estabelecida pela maneira como a estrutura de categorias foi organizada no banco de dados da Wikipédia. A Figura 16 apresenta um recorte da estrutura taxonômica obtida.

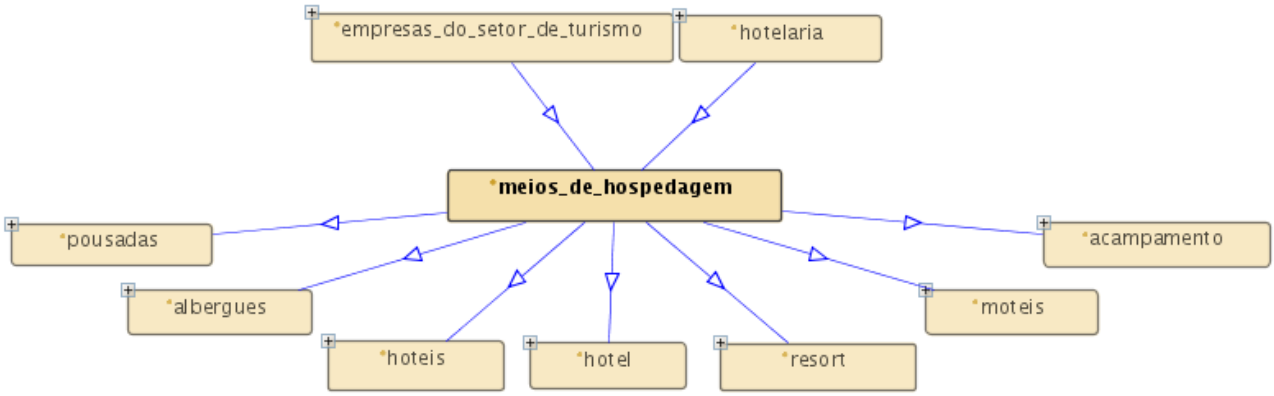


Figura 17 - Ligações do conceito “meios_de_hospedagem” na estrutura taxonômica extraída.

A estrutura taxonômica extraída apresenta-se como um grafo de relações semânticas e não como uma estrutura arbórea, conforme exemplo na Figura 17, em que a categoria “meios_de_hospedagem” aparece como subclasse das categorias “hotelaria” e “empresas_do_setor_de_turismo”.

5.1.2. Etapa 2: Identificação de Relações, Classes e Instâncias

Na segunda etapa é feita a análise dos conceitos presentes na taxonomia gerada na etapa anterior, e é realizada a extração de relações, novas classes e instâncias. Para

realizarmos esta tarefa foi necessário definir quais as relações e instâncias a serem extraídas dos títulos das classes da estrutura taxonômica, bem como as heurísticas utilizadas nesta extração.

Analisando a estrutura taxonômica extraída, percebemos que muitas das relações semânticas entre os conceitos não se caracterizavam pela relação de hiponímia (*is-a*) mas sim, pela relação *located-in*. Por exemplo, as duas categorias com maior número de subcategorias são “Transportes por país” e “Turismo por país”, ou seja, categorias cujo conteúdo semântico está relacionado com localização. Além disso, algumas categorias apresentam esta relação em seu título, como, por exemplo, “Termas do Brasil” que apresenta o relacionamento “termas” *located-in* “Brasil”. Deste modo, decidimos efetuar a extração de relações *located-in.*

A opção pela extração da relação *located-in* está diretamente ligada ao domínio representado, a partir da análise das subcategorias de Turismo na Wikipédia. Caso estivéssemos trabalhando com outro domínio, e logo, extraindo a estrutura ontológica a partir de outra categoria da enciclopédia, poderiamos ter optado por outras relações, como por exemplo, *part-of* ou *author-of*.

A partir da definição a respeito do uso da relação *located-in* na estrutura ontológica extraída, verificamos que esta relação não acontecia necessariamente entre classes, mas sim entre uma classe e uma instância de um lugar. No exemplo “termas” *located-in* “Brasil”, verificamos que “termas” é uma classe[24] e “Brasil” instância[25] de lugar.

Para realizar a tarefa de identificar os relacionamentos de localização e efetuar a distinção entre classes e instâncias, foram propostas quatro heurísticas que inferem a relação *located-in* a partir dos títulos dos conceitos da taxonomia gerada na primeira etapa do protótipo e distinguem classes e instâncias de local. A seguir, descrevemos estas heurísticas.

5.1.2.1. Heurística 1

Infere a existência de relacionamentos de localização em subclasses de classes cujo título contém as expressões “por país”, “por cidade” ou “por estado”.

Para cada classe da taxonomia, verificamos se o título contém as expressões “por país”, “por cidade” ou “por estado”. Em caso positivo, inferimos que todos os locais presentes no título das suas subcategorias possuem relação *located-in* com ela, conforme

[24] Neste estudo de caso consideramos classes como um conjunto de indivíduos.

[25] Neste estudo de caso consideramos instâncias como indivíduos que representam objetos únicos não passíveis de instanciação.

ilustrado pela Figura 18.

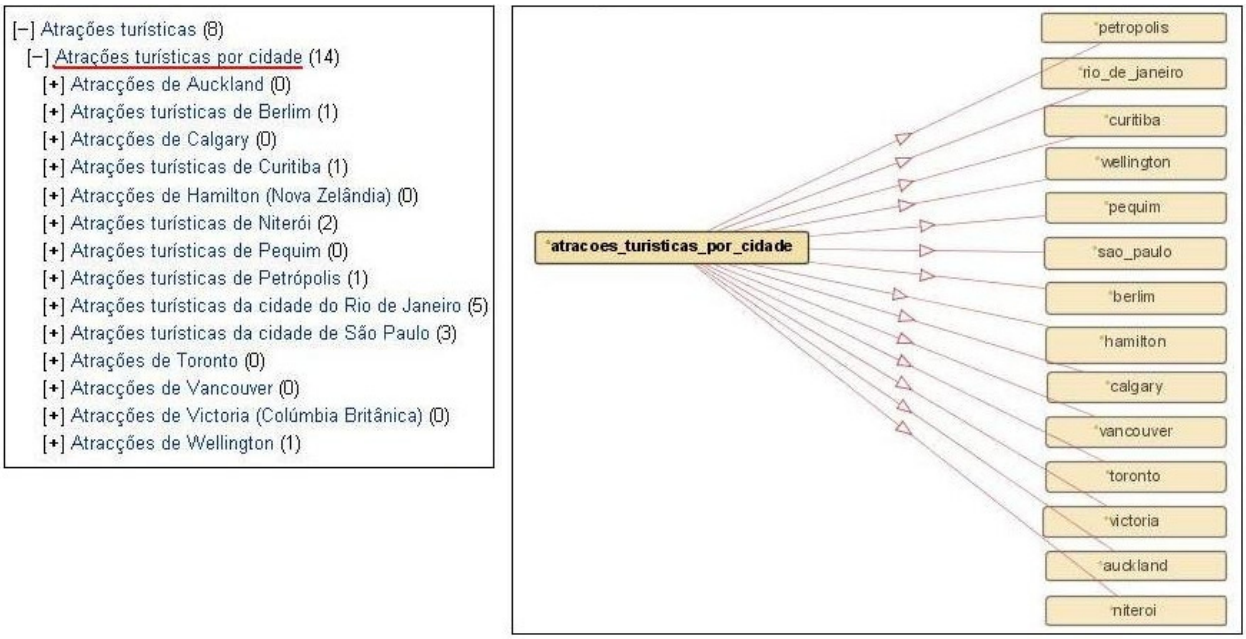


Figura 18 - Recorte da categoria “Atrações turísticas por cidade” na Wikipédia e representação da relação *located-in* com instâncias extraídas através da Heurística 1, após a execução do protótipo.

Exemplificando, “Atrações turísticas por cidade” contém “por cidade” em seu título, de tal forma que inferimos que todas suas subclasses apontam para locais, cabendo a inserção da relação *located-in* na estrutura ontológica.

Desejamos criar a relação “Atrações turísticas por cidade” *located-in* “Curitiba” porque “Atrações turísticas de Curitiba” é subclasse de “Atrações turísticas por cidade”, bem como criar a instância “Curitiba”. Para isso, seguimos os seguintes passos:

- Aplicamos a seguinte regra em relação ao título “Atrações Turísticas de Curitiba”: o nome da localidade (Curitiba) é a palavra que inicia em letra maiúscula, posicionada depois de uma preposição ou contração “de/do/da” e “em/no/na”.
- Criamos a instância “Curitiba” da classe “Local”.
- Excluímos a subclasse “Atrações turísticas de Curitiba” da taxonomia.
- Geramos a relação “Atrações turísticas por cidade” *located-in* “Curitiba”.

5.1.2.2. Heurística 2

Seu objetivo é inferir relacionamentos de localização em categorias contendo preposições ou contrações “de/do/da” e “em/no/na” em seu título, como, por exemplo, “Aeroportos da Argentina”.

Testamos, para cada classe da taxonomia, se contém em seu título as preposições

ou contrações “em/no/na” ou “de/do/da”. Caso positivo, iremos buscar se a palavra vizinha à preposição ou contração refere-se a um local e, neste caso, inferimos a existência de um relacionamento *located-in*.

Exemplificando: a categoria “Aeroportos da Argentina” possui a contração “da” em seu título e por isso inferimos que pode conter um local no seu nome, cabendo a inserção da relação *located-in* na estrutura ontológica. Para criar a relação “Aeroportos *located-in* Argentina” seguimos os seguintes passos:

- Verificamos se a palavra após a preposição ou contração identificada inicia por letra maiúscula. No caso, a palavra Argentina.
- Selecionamos todas as categorias relacionadas com “Argentina” e verificamos se alguma possui as palavras “município, província, cidade, estado, país ou reino” em seu título. Se possuir, concluímos que é uma localidade. No caso, “Argentina” possui ligação com uma categoria contendo “país” e criamos a instância “Argentina”.
- As palavras anteriores à localidade (Argentina), exceto a preposição ou contração anterior ao local identificado, são assumidas como uma nova classe, neste caso “Aeroportos”, que é criada.
- Excluímos a classe “Aeroportos da Argentina”.
- Criamos a relação “Aeroportos *located-in* Argentina”.

  A Figura 19 ilustra este exemplo.

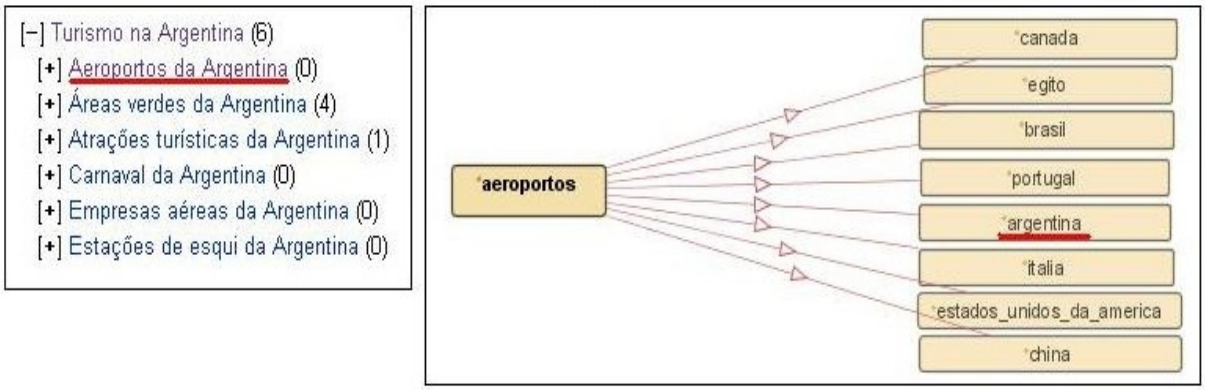


Figura 19 - Recorte da categoria “Aeroportos da Argentina” na Wikipédia e representação da relação *located-in* com as instâncias extraídas através da Heurística 2, após a execução do protótipo.

5.1.2.3. Heurística 3

Segundo esta heurística, classes contendo apenas uma palavra em seu título são candidatas a instâncias de lugar e possuem relação de localização com sua classe pai. Os passos para aplicação desta regra são os seguintes:

- Verificamos se a classe possui apenas uma palavra em seu título.
- Caso afirmativo, testamos se esta palavra refere-se a um local. Para isso, buscamos no banco de dados da Wikipédia se alguma das categorias ligadas à categoria que deu origem ao conceito possui as palavras “município, província, cidade, estado, país ou reino” em seu título. Se possuir, concluímos que a palavra descreve uma localidade.
- Excluímos a classe analisada e criamos uma instância de local com o mesmo título.
- Criamos um relacionamento *located-in* da classe pai da classe excluída, com a instância criada.

Exemplificando: “Cracovia” é subclasse de “Patrimônio Mundial da UNESCO”. Pesquisando as categorias com as quais “Cracovia” está conectada na Wikipédia, encontramos “Cidades da Polônia”. Excluímos a classe “Cracovia” e criamos a instância “Cracovia”. Por fim, criamos o relacionamento “Patrimônio Mundial da UNESCO” *located-in* “Cracovia”.

5.1.2.4. Heurística 4

Executada após o mapeamento de todas as classes e instâncias de acordo com as três heurísticas anteriores, sua meta é eliminar mapeamentos equivocados, de acordo com a seguinte regra:

- Se uma instância foi mapeada também como classe, mantém-se o mapeamento como classe e elimina-se o mapeamento como instância.

5.1.3. Etapa 3: Formatação e Unificação Linguística

Nesta etapa do protótipo, tomamos por entrada a estrutura ontológica contendo classes e instâncias grafadas do modo como estão cadastradas no banco de dados da Wikipédia e geramos a mesma estrutura ontológica com títulos formatados, permitindo sua descrição em OWL na próxima etapa.

Para isso, executamos uma função que:

- Substitui a sequência de caracteres “cç”, característica da grafia portuguesa, por “ç”.
- Remove acentos das palavras. Por exemplo, substituindo “ç” por “c”, “ã” por “a”, etc.
- Converte letras maiúsculas por minúsculas.
- Substitui espaços em branco por traço inferior (“_”).

Por exemplo, a classe “Áreas de proteção ambiental” tem sua grafia alterada para “areas_de_protecao_ambiental”.

5.1.4. Etapa 4: Geração da Descrição OWL

Na etapa final do protótipo criamos um arquivo contendo a descrição da estrutura ontológica na linguagem OWL.

O gerador de OWL foi implementado no protótipo em PHP, sem o auxílio de *frameworks* e outras ferramentas externas. O trecho de código OWL a seguir foi gerado pelo protótipo.

```
<owl:Class rdf:ID="termas">
<rdfs:subClassOf rdf:resource="#estancias_termais"/>
<rdfs:subClassOf>
        <owl:Restriction>
        <owl:onProperty rdf:resource="#located_in"/>
        <owl:someValuesFrom>
                <owl:Class>
                <owl:oneOf rdf:parseType="Collection">
                        <local rdf:ID="portugal"/>
                        <local rdf:ID="roma"/>
                        <local rdf:ID="brasil"/>
                </owl:oneOf>
                </owl:Class>
        </owl:someValuesFrom>
        </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="meios_de_hospedagem">
<rdfs:subClassOf rdf:resource="#hotelaria"/>
<rdfs:subClassOf rdf:resource="#empresas_do_setor_de_turismo"/>
</owl:Class>
```

A geração do código OWL é realizada seguindo os seguintes passos:

- Geração do cabeçalho OWL;
- Geração da classe local e suas instâncias;
- Geração da estrutura hierárquica de classes (relações *is-a*) e relações *located-in* entre classes e instâncias.

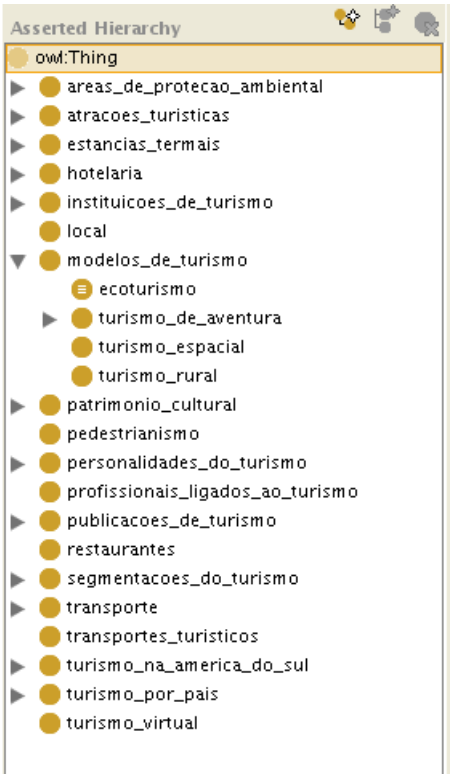


Figura 20 – Recorte da visualização das classes no Protégé.

A partir do arquivo OWL gerado é possível estender a ontologia manualmente ou com uso de ferramentas que apóiam a edição de ontologias. Como exemplo, na Figura 20 apresentamos parte do código OWL gerado pelo protótipo em um editor de ontologias[26].

5.1.5. Estrutura Ontológica Gerada

A estrutura ontológica criada após a execução do protótipo é composta por 165 classes e 156 instâncias. A Tabela 5 lista as instâncias presentes na estrutura e a Tabela 6 as classes hierarquicamente organizadas.

[26] Utilizamos a versão 3.4 do editor Protégé.

Tabela 5 – Instâncias da estrutura ontológica gerada pelo protótipo.

Instâncias			
africa	costa_rica	letonia	republica_checa
africa_do_sul	cracovia	liberia	republica_democratica_do_congo
albania	croacia	luxemburgo	republica_dominicana
alemanha	curitiba	malasia	rio_de_janeiro
america_do_norte	dakota_do_sul	malawi	roma
america_do_sul	dinamarca	marrocos	romenia
angola	egito	massachusetts	russia
antilhas_neerlandesas	emirados_arabes_unidos	mexico	sao_paulo
arabia_saudita	equador	mianmar	senegal
argelia	escocia	michigan	serra_leoa
argentina	eslovaquia	miradouros	servia
armenia	eslovenia	mocambique	sudao
asia	espanha	moldavia	suecia
auckland	estados_unidos_da_america	mongolia	suica
australia	estonia	montenegro	supervia
austria	europa	namibia	suriname
azerbaijao	filipinas	nepal	tailandia
bahrein	finlandia	niger	texas
belgica	florida	nigeria	tiradentes_minas_gerais
benim	franca	niteroi	toronto
berlim	georgia	noruega	turquia
bielorrussia	grecia	nova_iorque	uruguai
bolivia	guine	nova_jersei	utah
brasil	hamilton_nova_zelandia	nova_zelandia	uzbequistao
bulgaria	honduras	oceania	vancouver
burkina_faso	hungria	oma	venezuela
cabo_verde	india	paises_baixos	victoria_columbia_britanica
calgary	indonesia	panama	vietna
california	inglaterra	paquistao	vietname
camaroes	ira	paraguai	virginia
camboja	iraque	paris	vitoria
canada	irlanda	pensilvania	wellington
cartago	irlanda_do_norte	pequim	zimbabwe
casinos	islandia	peru	
cazaquistao	israel	petropolis	
checoslovaquia	italia	polonia	
chile	jamaica	porto_rico	
china	japao	portugal	
colombia	jerusalem	puebla	
coreia_do_norte	jordania	quenia	
costa_do_marfim	kuwait	reino_unido	

Tabela 6 - Classes da estrutura ontológica gerada pelo protótipo.

- turismo_virtual
  - patrimonio_cultural
    - registro_nacional_de_lugares_historicos
      - lugares_historicos_registrados_no_distrito_de_columbia
      - lugares_historicos_registrados
    - patrimonio_cultural_imaterial
    - sitios_arqueologicos
      - cidades_da_antiguidade
      - sitios_arqueologicos_da_mesoamerica
      - petra
    - monumentos
      - sete_maravilhas_do_mundo
      - monumentos_religiosos
      - obeliscos
      - monumentos_da_roma_antiga
      - monumentos_funerarios
      - monumentos_naturais
      - monumentos_megaliticos
      - monumentos_comemorativos
    - listas_de_patrimonio
    - patrimonio_cultural_por_pais
    - patrimonio_mundial_da_unesco
      - amazonia
      - nubia
  - transporte
    - infra-estrutura_de_transportes
      - infra-estrutura_aeroportuaria
      - infra-estrutura_ferroviaria
      - infra-estrutura_rodoviaria
      - infra-estrutura_hidroviaria
    - logistica
      - empresas_de_logistica
      - gasodutos
      - transporte_hidroviario
      - transporte_rodoviario
      - transporte_tubular
      - transporte_ferroviario
      - transporte_aereo
    - acidentes_de_transporte
      - acidentes_ferroviarios
      - acidentes_aereos
      - acidentes_de_transito
    - aviacao
      - terminologia_da_aviacao
      - tipos_de_aviacao
      - seguranca_aerea
      - veiculos_aereos_nao_tripulados

- aeroclubes
- profissionais_ligados_a_aviacao
- historia_da_aviacao
- organizacoes_aeronauticas
- controle_de_trafego_aereo
- handling
- forca_aerea
- avionica
- avioes
- escolas_de_aviacao
- aviadores
- helicopteros
- aeromodelismo
- sistemas_de_reservas
- (transporte_aereo)...
- aeroportos
- empresas_aereas

transito
transportadoras
transportes
meios_de_transporte
- taxi
- veiculos
- jet_ski
- (transporte_hidroviario)...
- (transporte_rodoviario)...
- (transporte_tubular)...
- (transporte_ferroviario)...
- (transporte_aereo)...

transportes_publicos
- central
- vlt
- transporte_publico
- companhia_paulista_de_trens_metropolitanos
- empresas_de_transportes_publicos
- Ônibus
- metro
- metropolitanos
- expresso_tiradentes

transportes_por_continente
transportes_por_pais
- acidentes_de_transporte_por_pais
- transporte_ferroviario_por_pais
- transporte_hidroviario_por_pais
- transporte_rodoviario_por_pais

prefeitos
turismo_na_costa_rica
atracoes_turisticas

```
      parques_tematicos
        parques_tematicos_por_pais
      casinos
      atracoes_turisticas_por_cidade
      estacoes_de_esqui
      atracoes_turisticas_por_pais
    parques_nacionais
estancias_termais
    termas
personalidades_do_turismo
    politicos_do_turismo
      ministros_do_turismo
profissionais_ligados_ao_turismo
publicacoes_de_turismo
    jornais_de_turismo
    revistas_de_turismo
hotelaria
    meios_de_hospedagem
      moteis
      resort
        (estacoes_de_esqui)...
      albergues
      hotel
      acampamento
      pousadas
      hoteis
        redes_hoteleiras
        hoteleiros
instituicoes_de_turismo
    fundacoes_de_turismo
    empresas_do_setor_de_turismo
      cafes_e_restaurantes_do_grande_porto
      empresas_de_cruzeiros
      royal_caribbean_international
      (meios_de_hospedagem)...
areas_de_protecao_ambiental
    natura_2000
    ilha_comprida
    sitios_ramsar
    parques_estaduais
    reservas_da_biosfera
      reservas_da_biosfera_na_america_latina_e_no_caribe
    (parques_nacionais)...
local
modelos_de_turismo
    ecoturismo
      tres_coroas
    turismo_de_aventura
```

```
        esporte_de_aventura
    turismo_rural
    turismo_espacial
turismo_na_america_do_sul
    conservacao
    areas_verdes
    patrimonio_historico
    centros_de_convencoes
    patrimonio_edificado
    carnaval
    restaurantes
    turismo
    (aeroportos)...
    (patrimonio_mundial_da_unesco)...
    (empresas_aereas)...
    (hoteis)...
    (estacoes_de_esqui)...
transportes_turisticos
turismo_por_pais
    regioes_turisticas
    jardins_zoologicos_nos_estados_unidos_da_america
    jardins_zoologicos
    bares
    rotas_dos_vinhos
    turismo_em_portugal
        regioes_de_turismo
        miradouros
        (patrimonio_edificado)...
        praias
        (termas)...
        (pousadas)...
    marinas
    (monumentos)...
    (conservacao)...
    (areas_verdes)...
    (patrimonio_historico)...
    (centros_de_convencoes)...
    (patrimonio_edificado)...
    (carnaval)...
    (restaurantes)...
    (praias)...
    (turismo)...
    jardins_zoologicos_por_pais
    (aeroportos)...
    (patrimonio_mundial_da_unesco)...
    (empresas_aereas)...
    (hoteis)...
    (estacoes_de_esqui)...
```

```
    (atracoes_turisticas)...
  segmentacoes_do_turismo
    turismo_de_caca
    turismo_cultural
    turismo_de_eventos
      (centros_de_convencoes)...
    turismo_de_negocios
    (ecoturismo)...
    (turismo_de_aventura)...
    (turismo_rural)...
    (turismo_espacial)...
  pedestrianismo
```

A seguir relatamos a avaliação da estrutura ontológica gerada por este estudo de caso e discutimos os resultados alcançados.

6. AVALIAÇÃO DOS RESULTADOS

Com o intuito de avaliar os resultados obtidos com o estudo de caso relatado no capítulo anterior, buscamos examinar em detalhe estes resultados. A estrutura ontológica gerada foi avaliada através da comparação com uma estrutura de referência (G*olden Mapping*) construída manualmente, revisada e refinada por uma linguista. Foram calculadas as medidas de Precisão, Abrangência e Medida-F, que são discutidas ao longo do capítulo.

6.1. Metodologias de Avaliação

Procuramos metodologias que permitissem examinar o desempenho da técnica de extração proposta, através da avaliação da estrutura ontológica obtida na execução do protótipo. Optamos, fundamentados no trabalho de Euzenat [EUZ07], por métricas baseadas em um modelo ideal, ou seja, uma ontologia de referência (*Golden Mapping*), para avaliar a estrutura ontológica gerada após a execução do protótipo.

[PON07b, ZIR08, NAS08] avaliam seus resultados através da comparação com os dados presentes na base de dados ResearchCyc e na WordNet. Visto que não foram encontradas, até o presente momento, bases de dados em português com as quais fosse possível realizar a comparação dos resultados obtidos neste trabalho, optamos por elaborar manualmente a estrutura ontológica de referência.

Para isso, tomamos por base a estrutura taxonômica da categoria Turismo da Wikipédia em três níveis, e a partir dela classificamos manualmente as novas classes e instâncias em um editor de ontologias. Em seguida, esta estrutura foi avaliada e refinada por uma linguista. A estrutura gerada foi utilizada como o *Golden Map* na obtenção das métricas de Precisão, Abrangência e Medida-F.

A Precisão mede a taxa de correspondência correta em relação ao total de correspondências obtidas. A Abrangência mede o número de correspondências obtidas em relação ao total de correspondências esperadas. A Figura 21 representa estes conceitos tal como a fonte citada. Também é empregada a Medida-F (média harmônica entre Precisão e Abrangência).

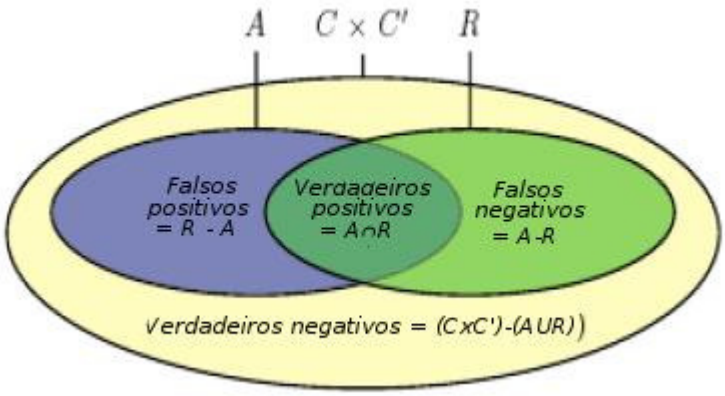


Figura 21 - Representação das medidas de Precisão e Abrangência - adaptação de [EUZ07].

6.2. Definição (Precisão, Abrangência e Medida-F)

Dado um alinhamento de referência R, a Precisão P de um alinhamento A é dada por

$$P(A,R) = \frac{|R \cap A|}{|A|}$$

A Abrangência Ab é dada por

$$Ab(A,R) = \frac{|R \cap A|}{|R|}$$

A Medida-F é dada pelo cálculo da média harmônica entre as duas medidas (Precisão e Abrangência)

$$F = \frac{2.(P \,.\, Ab)}{P + Ab}$$

6.3.Processo de Avaliação e Resultados Obtidos

Para avaliar a qualidade da estrutura ontológica gerada pelo protótipo implementado, buscamos medidas baseadas na comparação entre resultados esperados e resultados obtidos. Decidimos utilizar as medidas Precisão, Abrangência e Medida-F, usuais em extração da informação, e adaptadas para ontologias [EUZ07].

Nesta avaliação, o Alinhamento (*A)* consiste na estrutura ontológica gerada pelo protótipo, em que utilizamos a metodologia de extração de estruturas ontológicas da Wikipédia proposta nesta dissertação. A Referência *(R)* é uma estrutura ontológica elaborada manualmente a partir da estrutura da categoria Turismo da Wikipédia em língua portuguesa, revisada e refinada por uma linguista.

A Tabela 7 apresenta o tamanho das duas estruturas ontológicas (alinhamento e referência), listando o número de classes presentes nas estruturas ontológicas *A* e *R*.

Tabela 7 - Número de classes das estruturas ontológicas de Alinhamento e Referência.

Estrutura Ontológica	Número de Classes	Número de Instâncias
Alinhamento	165	163
Referência	144	156

Realizamos a avaliação do protótipo em relação à identificação de instâncias da classe local, identificação de relações *located-in,* identificação de relações *is-a* e a estrutura ontológica completa.

A seguir tecemos algumas considerações a respeito da construção da estrutura ontológica de referência.

6.3.1. Construção do Modelo de Referência

O modelo de Referência *(R)* é uma estrutura ontológica elaborada manualmente a partir da estrutura da categoria Turismo da Wikipédia em língua portuguesa, revisada e refinada por uma linguista, refletindo sua interpretação da caracterização deste domínio. O relatório a respeito da revisão e refinamento desta estrutura, elaborado pela linguista convidada, constitui interessante ferramenta para discussão dos caminhos que podem ser trilhados, e encontra-se no Anexo B desta dissertação.

O envolvimento da linguista na construção do modelo de referência foi motivado pela necessidade de obter-se uma avaliação isenta do estudo de caso realizado. Desta forma, a estrutura ontológica utilizada como *Golden Mapping* mantém o olhar da linguista para garantir a imparcialidade dos resultados obtidos.

A estrutura de referência foi elaborada adotando os seguintes passos:

1. Exportação da estrutura da categoria Turismo da Wikipédia em Português em três níveis para uma taxonomia descrita em OWL.
2. Construção manual[27], a partir da estrutura taxonômica gerada no passo anterior, de uma estrutura ontológica do domínio Turismo contendo classes, instâncias da classe Local e relações *is-a* e *located-in*.
3. Revisão e refinamento da estrutura ontológica gerada no segundo passo por uma linguista, conforme relatório incluído no Anexo B.

A seguir apresentamos e discutimos os resultados obtidos na avaliação.

[27] A manipulação foi feita com o editor de ontologias Protégé 3.4.

6.4. Resultados Obtidos

6.4.1. Instâncias da Classe Local

Tabela 8 - Avaliação do mapeamento das instâncias de Local.

Métrica	Resultado
Precisão	0.993506493506
Abrangência	0.944444444444
Medida-F	0.96835443038

A partir da medida de Precisão obtida (Tabela 8), procuramos identificar os locais mapeados na estrutura de referência e não mapeados pelo protótipo, e investigar os motivos das falhas.

Observa-se que apenas “supervia” foi mapeado como local na estrutura ontológica gerada pelo protótipo e não foi encontrada na referência, tratando-se de um falso positivo.

Esta falha ocorreu na terceira heurística, descrita na seção 5.1.2.3 deste documento. A classe “supervia” da estrutura taxonômica gerada na primeira etapa do protótipo, contém apenas uma palavra em seu título. Sua categoria correspondente no banco de dados da Wikipédia está conectada à categoria “Transportes da cidade do Rio de Janeiro” que contém a palavra “cidade” em seu título. Visto que a heurística 3 define que classes com apenas uma palavra ligadas a categorias contendo a palavra cidade são mapeadas como local, este é o motivo do mapeamento equivocado.

Ao analisar os resultados da Abrangência verificamos que, no total, nove instâncias da classe “local”, mapeadas na estrutura de referência, não foram mapeadas na estrutura gerada pelo protótipo: grande_porto, columbia, petra, mesoamerica, tres_coroas, amazonia, america_latina, nubia e ilha_comprida.

A principal causa da falha neste mapeamento é a ausência de ligações, no banco de dados, entre as categorias correspondentes às classes que deveriam ter sido mapeadas como locais, com outras categorias ligadas à localização. Esta foi a causa do não mapeamento, como instâncias de “local”, das classes “columbia”, “mesoamerica”, “amazonia”, “nubia”, “cafes_e_restaurantes_do_grande_porto” e “petra”.

6.4.2. Relações *Located-in*

Tabela 9 - Avaliação do mapeamento das relações *located-in.*

Métrica	**Resultado**
Precisão	0.841648590022
Abrangência	0.919431279621
Medida-F	0.878822197055

Embora os resultados apresentados na Tabela 9 pareçam bastante positivos, em relação à literatura, para compreender estes números, buscamos identificar os mapeamentos realizados na estrutura de referência e não mapeados pelo protótipo, e investigar os motivos das falhas.

Foram encontradas duas falhas principais. A primeira, está relacionada ao não mapeamento de classes como instância de Local, o que já foi analisado na seção anterior. A segunda, está relacionada a casos como o seguinte: “turismo_na_argentina” e “turismo_no_brasil” são subclasses de “turismo_na_america_do_sul”, e caíram na segunda heurística, descrita na seção 5.1.2.2, sendo mapeadas como “turismo” *located-in* “argentina” e “turismo” *located-in* “brasil”. Entretanto, na estrutura de referência elas estão mapeadas como “turismo_na_america_do_sul” *located-in* “argentina” e “turismo_na_america_do_sul” *located-in* “brasil”.

A lista completa das relações *located-in* mapeadas pela referência, mas não pelo protótipo e também daquelas relações que foram mapeadas pelo protótipo, mas não pela referência, encontra-se no Anexo A da dissertação.

6.4.3. Relações *Is-a*

Tabela 10 - Avaliação do mapeamento das relações *is-a.*

Métrica	**Resultado**
Precisão	0.730303030303
Abrangência	0.919847328244
Medida-F	0.814189189189

A partir da medida de Abrangência (Tabela 10), identificamos que 13 relações de subsunção foram mapeadas na estrutura de referência e não pelo protótipo. Estas falhas foram causadas, principalmente pela Heurística 1 (seção 5.1.2.1), onde instâncias de

Local são extraídas das subclasses de classes contendo “por país”, “por cidade” ou “por estado” em seu título. Por exemplo, “patrimonio_edificado_do_peru” é subclasse de “patrimonio_cultural_por_pais”. Aplicando a Heurística 1, “peru” torna-se instância de local e cria-se a relação “patrimonio_cultural_por_pais” *located_in* “peru”, havendo a exclusão de “patrimonio_edificado”, que é uma classe da estrutura ontológica de referência.

Em relação à Precisão, o principal fato gerador do mapeamento equivocado de relações *is-a* pelo protótipo foi a aplicação da Heurística 2 (Seção 5.1.2.2). Esta regra gera relações *located-*in em classes contendo em seu título as preposições ou contrações “em/no/na” ou “de/do/da”. Nesta tarefa é feita uma decomposição do título classe, de onde se extrai uma nova classe e uma instância de Local.

Por exemplo, na estrutura ontológica de referência, a classe “hoteis” é subclasse de “meios_de_hospedagem”, enquanto que na estrutura avaliada, a classe “hoteis” também é subclasse de “turismo_na_america_do_sul”, visto que o protótipo extrai da classe “hoteis_do_brasil”, a classe “hoteis” e a instância “brasil”, para criar a relação “hoteis” *located-in* “brasil” e posiciona a classe “hoteis” como subclasse de “turismo_na_america_do_sul”, posição original da classe “hoteis_do_brasil”.

Outro fator gerador de diferença entre o mapeamento de relações *is-a* presentes na estrutura ontológica de referência e na estrutura avaliada foram algumas divergências entre a estrutura hierárquica das categorias criada pelos colaboradores da Wikipédia e na hierarquia proposta na estrutura de referência. Exemplificando, a classe “Cidades da antiguidade” foi mapeada no banco de dados da Wikipédia, e consequentemente na estrutura ontológica gerada pelo protótipo, como subclasse de “Sítios Arqueológicos”. Entretanto, a autora da referência posicionou a classe “Cidades da antiguidade” como subclasse de “Atrações turísticas”.

A Tabela 3, que se encontra na Seção 3.2.1, apresenta os resultados da avaliação das relações *is-a* do trabalho descrito em [PON07b]. Esta avaliação foi realizada comparando os resultados com os pares gerados contendo conceitos correspondentes na ontologia ResearchCyc (85% dos pares gerados pelos autores)[7]. O resultado foi uma Precisão de 89,1% e Abrangência de 86,6%.

Verificamos que nosso trabalho obteve melhores resultados em relação à Abrangência, mas um número menos significativo quanto à Precisão. Tal resultado nos mostra que é preciso buscar maneiras mais eficientes para inferência das relações de subsunção, que somente a estrutura de categorias da Wikipédia.

A lista completa das relações *is-a* mapeadas pela referência, mas não pelo protótipo, e também daquelas que foram mapeadas pelo protótipo, mas não pela

referência, encontra-se no Anexo A da dissertação.

6.4.4. Estrutura Ontológica Completa

A Tabela 11 apresenta a aferição das métricas da estrutura ontológica gerada pelo protótipo como um todo, avaliando classes, instâncias e relações *is-a* e *located-in.* Ela condensa os dados apresentados anteriormente nas Tabelas 8, 9 e 10, em que avaliamos separadamente os diferentes tipos de informação da estrutura ontológica extraída.

Tabela 11 - Avaliação da estrutura ontológica gerada pelo protótipo.

Métrica	**Resultado**
Precisão	0.795195954488
Abrangência	0.919590643275
Medida-F	0.852881355932

As métricas apresentadas na Tabela 11 mostram que a estrutura ontológica gerada pelo protótipo aproxima-se bastante da estrutura de referência, o que demonstra a viabilidade da extração de estruturas ontológicas de domínio em português a partir das categorias da Wikipédia através do método proposto.

As principais diferenças entre a estrutura ontológica gerada pelo protótipo e a referência encontram-se no mapeamento da relação *is-a*. As causas destas diferenças foram discutidas na seção 6.4.3. O protótipo desenvolvido no estudo de caso teve seu melhor desempenho no mapeamento das instâncias da classe “Local”, como detalhado na Seção 6.4.1, com Precisão próxima aos 100%.

A partir destas considerações, acreditamos que, em relação ao método proposto, a definição de heurísticas adequadas é um ponto chave para que a extração seja bem sucedida, gerando uma estrutura ontológica que retrate adequadamente o domínio descrito.

A seguir apresentaremos as considerações finais quanto ao trabalho realizado, descrevendo suas principais contribuições e destacaremos os rumos para futuras pesquisas na área.

7. CONSIDERAÇÕES FINAIS

Esta dissertação buscou contribuir para a solução do problema da carência de ontologias disponíveis, particularmente em língua portuguesa. Nela foi proposta um método semi-automático para extração de estruturas ontológicas de domínio a partir da estrutura de categorias da Wikipédia em português.

Para realizar esta pesquisa foi feito um estudo da base teórica a respeito de ontologias e das iniciativas propondo a extração de estruturas ontológicas da Wikipédia. Em seguida, identificamos técnicas a serem utilizadas para aquisição de conhecimento a partir da Wikipédia em português e definimos uma abordagem para a extração de estruturas ontológicas a partir desta versão da enciclopédia.

A partir destes estudos e definição, foi proposto um método semi-automático de extração de estruturas ontológicas de domínio a partir das categorias da Wikipédia em língua portuguesa, avaliado através de um estudo de caso que produziu uma estrutura ontológica do domínio Turismo contendo classes, instâncias e relações. O protótipo implementado no estudo de caso gerou uma estrutura ontológica descrita em OWL avaliada através da comparação automática com uma estrutura de referência (*Golden Mapping*) construída manualmente, revisada e refinada por uma linguista. Foram calculadas as medidas de Precisão, Abrangência e Medida-F para avaliar os resultados obtidos.

Os resultados apresentados no Capítulo 6 confirmam a viabilidade do método semi-automático de extração de estruturas ontológicas de domínio da Wikipédia, proposto nesta dissertação. Nosso método de extração a partir da estrutura de categorias da enciclopédia é, a nosso ver, bastante promissor, particularmente na etapa de extração do processo de construção de ontologias, apresentado na seção 2.1 deste documento.

7.2. Discussão Sobre os Resultados

Para avaliar a eficiência do método proposto, medimos Precisão, Abrangência e Medida-F do mapeamento de instâncias da classe Local, do mapeamento de relações *located-in* e *is-a* e da estrutura ontológica completa. A partir destes resultados, investigamos algumas das causas de acertos e erros do protótipo, buscando examinar as similaridades e discrepâncias entre a estrutura ontológica de referência e a estrutura gerada no estudo de caso.

As principais diferenças entre a estrutura ontológica gerada pelo protótipo e a referência encontram-se no mapeamento da relação *is-a*. Percebemos que um dos

motivos que geraram esta desigualdade foi o fato da exclusão das repetições das classes ser um dos princípios da construção humana do *Golden Map*, conforme relatório no Anexo B. Além disso, outras causas destas diferenças foram discutidas na Seção 6.4.3. O protótipo desenvolvido no estudo de caso teve seu melhor desempenho no mapeamento das instâncias da classe “Local”, detalhado na Seção 6.4.1.

A partir dos erros e acertos cometidos durante a implementação do protótipo, acreditamos que a definição de heurísticas adequadas é o ponto chave para que a extração seja bem sucedida, gerando uma estrutura ontológica que retrate adequadamente o domínio descrito.

Consideramos os resultados satisfatórios: 79.51% de Precisão e 91.95% de Abrangência, considerando a estrutura ontológica como um todo. Estes resultados mostram que a extração de estruturas ontológicas da Wikipédia é tema que oferece boas perspectivas de aprofundamentos.

O método de avaliação utilizado, bem como a elaboração manual do *Golden Map* e revisão por uma linguista, mostraram-se adequados para este trabalho. Entretanto, em razão do caráter subjetivo da modelagem de ontologias, é possível discutir se a abordagem da linguista, evitando a repetição das classes foi adequada. Outras opções de avaliação podem ser buscadas, como a revisão da estrutura ontológica de referência por um cientista da informação e o uso de ontologias criadas a partir de outras fontes, por exemplo. Também seria interessante a realização de uma avaliação extrínseca, através por um sistema de recuperação de informação, por exemplo.

7.3. Contribuições deste Estudo

Um método confirmando a viabilidade do uso da estrutura de categorias da Wikipédia em língua portuguesa na extração de estruturas ontológicas, constituiu a principal contribuição deste trabalho. Em uma visão mais detalhada, as contribuições alcançadas foram:

- Levantamento e análise de abordagens e técnicas para extração de estruturas ontológicas da Wikipédia;
- Método semi-automático para extração de estruturas ontológicas das categorias da Wikipédia em língua portuguesa;
- Estudo de caso validando o método proposto, apoiado por um protótipo desenvolvido como uma contribuição prática;
- Estrutura ontológica do domínio Turismo em português extraída da Wikipédia, que poderá ser empregada, por exemplo, no enriquecimento de outras ontologias ou

como ponto de partida de outros trabalhos;

- Avaliação da estrutura ontológica obtida.

7.4. Publicações

Os artigos relacionados a seguir foram produzidos e publicados durante o período do mestrado e seu conteúdo se refere ao trabalho conduzido:

- Construção de uma Estrutura Ontológica de Domínio a partir da Wikipédia [XAV09a]

  Publicado nos anais e apresentado na seção principal do STIL2009 (7th Brazilian Symposium in Information and Human Language Technology). Este artigo apresenta resultados parciais do trabalho realizado em nosso mestrado, relatando um estudo sobre a extração de uma estrutura ontológica contendo relações de hiponímia e localização, da categoria Turismo da Wikipédia em português, ilustrado por um experimento, com a avaliação dos resultados obtidos.

- Extração de Estruturas Ontológicas a partir da Wikipédia em Língua Portuguesa [XAV09b]

  Publicado nos anais do Webmedia 2009 (Simpósio Brasileiro de Sistemas Multimídia e Web) e apresentado no WTD (IX Workshop de Teses e Dissertações). Este artigo relata o andamento de nossa dissertação até setembro de 2009, apresentando o tema central de nossa pesquisa, seus objetivos, fundamentação teórica e resultados parciais.

7.5. Trabalhos Futuros

Pretendemos que este trabalho tenha sua continuidade no doutorado a ser iniciado no primeiro semestre de 2010 nesta mesma instituição. Nosso objetivo é aprimorar o método de extração de estruturas ontológicas da Wikipédia em língua portuguesa proposto no mestrado, buscando sua automatização.

Para realizar esta tarefa, apontamos o uso de outros elementos da enciclopédia, como *infoboxes*, texto dos artigos e a ligação entre artigos, além da estrutura de categorias, como fonte de dados. Para automatizar, são necessárias estratégias para a inferência de relações e extração de classes e instâncias, que podem ser alcançadas através de ações como:

- Mecanismo para automatizar a delimitação da profundidade da busca na primeira etapa do método.
- Método para inferência de relações entre termos de uma estrutura ontológica, a

partir da Wikipédia, independente de domínio.

- Método para extração de classes e instâncias da Wikipédia, independente de domínio.
- Mecanismo para avaliação de estruturas ontológicas a ser utilizado no desenvolvimento do projeto.

REFERÊNCIAS

[AUE07a] Auer, S.; Lehmann, J. “What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content”. In: European Semantic Web Conference (ESWC'07), 2007, pp.503-517.

[AUE07b] Auer, S.; Bizer C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z.G. “Dbpedia: A nucleous for a web of open data”. In: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007), vol. 4825, Chapter 52, Springer, 2007, pp.722-735.

[BER99] Berners-Lee, T.; Hendler, J.; Lassil, O. “The semantic web”. *Scientific American*, vol. 284 - 5, 2001, pp.34-43.

[BRA00] Brants, T. “TnT – A statistical Part-of-Speech tagger”. In: 6th Applied Natural Language Processing Conference (ANLP-2000), 2000, pp.224–231.

[BRE05] Breitman, K. K. “Web semântica**:** a internet do futuro”. Rio de Janeiro: LTC, 2005, 190 p.

[BUF08] Buffa, M.; Gandon, F.; Ereteo, G.; Sander, P.; Faron, C. “SweetWiki: A semantic wiki”. *Web Semantics*, vol. 6-1, fevereiro 2008, pp.84-97.

[BUI05] Buitelaar, P.; Cimiano, P.; Magnini, B. “Ontology Learning from Text: Methods, Evaluation and Applications”. IOS Press, 2005,180 p.

[CAR99] Caraballo, S. A. “Automatic construction of a hypernym-labeled noun hierarchy from text”. In: 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999), 1999, pp.120-126.

[CIM06a] Cimiano, P. “Ontology Learning and Population from Text: Algorithms, Evaluation and Applications”. Springer, 2006, 347 p.

[DAV06] Davies, J.; Studer, R.; Warren, P. “Semantic Web Technologies: Trends and Research in Ontology-Based Systems”. John Wiley & Sons, 2006, 326 p.

[EUZ07] Euzenat, J. “Semantic precision and recall for ontology alignment evaluation”. In: 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp.348-353.

[FIN05] Finkel, J. R.; Grenager, T.; Manning, C. “Incorporating non-local information into information extraction systems by Gibbs sampling”. In: 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 2005,

pp.363-370.

[GIR06] Girju, R.; Badulescu, A.; Moldovan, D. “Automatic discovery of part-whole relations”. *Computational Linguistics*, vol. 32 -1, 2006, pp.83-135.

[GRU92] Gruber, T. R. “Ontolingua: A Mechanism to Support Portable Ontologies”. Relatório Técnico, 1992, 61 p.

[GRU93] Gruber, T. R. “Towards Principles for the Design of Ontologies Used for Knowledge Sharing”. *International Journal of Human and Computer Studies*, vol. 43–5–6, 1993, pp.907-928.

[GUA98] Guarino, N. “Formal ontology and information systems”. In: 1st. Conference on Formal Ontology in Information Systems (FOIS'98), 1998, pp.3-15.

[HAR93] Harman, D.; Liberman, M. “TIPSTER Complete”. Linguistic Data Consortium, 1993. DVD.

[HEA92] Hearst, M. A. “Automatic acquisition of hyponyms from large text corpora”. In: 14th International Conference on Computational Linguistics, Nantes (COLING 1992), 1992, pp.539-545.

[HEP06] Hepp, M.; Bachlechner, D.; Siorpaes, K. “Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements”. In: Workshop on Semantic Wikis at the 3rd Annual European Semantic Web Conference (SemWiki 2006 at ESCW’06), pp.54-65.

[HOR04] Horridge, C.; Knublauch, H.; Rector, A.; Stevens, R.; Wroe, C. “A practical guide to building OWL Ontologies using the Protégé OWL Plug-in and CO-ODE Tools”. Disponível em: http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf, agosto 2004. Acesso em: janeiro 2010.

[KIT08] Kittur, A.; Kraut, R. E. “Harnessing the wisdom of crowds in wikipedia: quality through coordination”. In: ACM Conference on Computer-Supported Cooperative Work (CSCW 2008), 2008, pp.37-46

[KRÖ05] Krötzsch, M.; Vrandecic, D.; Völkel, M. “Wikipedia and semantic web - the missing links”. In: 1st International Wikimedia Conference (Wikimania 2005), Frankfurt, Alemanha, 2005. Anais eletrônicos. Disponível em: http://meta.wikimedia.org/wiki/Transwiki:Wikimania05. Acesso em: janeiro 2010.

[KUD00] Kudoh, T.; Matsumoto, Y. “Use of Support Vector Machines for chunk

identification". In: 4th Conference on Computational Natural Language Learning (CoNLL-2000), 2000, pp.142-144.

[LAU09] Lau, R.Y.K.; Dawei, S.; Yuefeng, L.; Cheung, T.C.H.; Jin-Xing, H. "Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning". *IEEE Transactions on Knowledge and Data Engineering*, vol. 21-6, junho 2009, pp.800–813.

[LIM07] Lima, V. L. S.; Nunes, M. G. V.; Vieira, R. "Desafios do Processamento de Línguas Naturais". In: 34º Seminário Integrado de Software e Hardware (SEMISH 2007), 2007, pp.2202-2216.

[MAE01] Maedche, A.; Staab, S. "Ontology Learning for the Semantic Web". *IEEE Intelligent Systems*, vol. 16, 2001, pp.72-79.

[MAE02] Maedche, A. "Ontology Learning for the Semantic Web". Kluwer Academic Publishers, 2002, 244p.

[MCG04] McGuinness, D. L.; Harmelen, F. "OWL Web Ontology Language Overview". Disponível em: http://www.w3.org/TR/owl-features/, fevereiro 2004. Acesso em: janeiro 2010.

[MIL95] Miller, G. A. "WordNet: A lexical database for English". *Communications of the ACM*, vol 38-11, novembro 1995, pp.39-41.

[MIL06] Miller, G. A.; Hristea, F. "WordNet nouns: Classes and instances". *Computational Linguistics*, vol. 32-1, março 2006, pp.1-3.

[MIK08] Mika, P.; Ciaramita, M.; Zaragoza, H.; Atserias, J. "Learning to Tag and Tagging to Learn: A Case Study on Wikipedia". *IEEE Intelligent Systems*, vol. 23-5, setembro 2008, pp.26-33.

[NAS08] Nastase, V; Strube, M. "Decoding Wikipedia Categories for Knowledge Acquisition". In: 23rd AAAI Conference on Artificial Intelligence (AAAI-08), 2008, pp.1219-1224.

[NOY01] Noy, N. F.; McGuinness, D. L. "Ontology Development 101: A Guide to Creating Your First Ontology". Disponível em: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf, março 2001. Acesso em: janeiro 2010.

[PAN07] Pan J. Z. "OWL for the Novice - A Logical Perspective". In: Baker, C. J. O.; Cheung K. (Ed.). *Semantic Web: Revolutionizing Knowledge Discovery in the*

Life Sciences. Springer, 2007, pp.159-182.

[PER99] Pérez, A. G. “Ontological engineering: A state of the art”. In: *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence,* vol. 2-3, 1999, pp.33-43.

[PON07a] Ponzetto, S. P.; Strube, M. “Knowledge derived from wikipedia for computing semantic relatedness”. *Journal of Artificial Intelligence Research,* vol. 30, 2007, pp.181-212.

[PON07b] Ponzetto S. P.; Strube M. “Deriving a large scale taxonomy from Wikipedia”. In: 22nd Conference on Artificial Intelligence (AAAI-07), 2007, pp.1440-1445.

[PON08] Ponzetto S. P.; Strube M. “WikiTaxonomy: A Large Scale Knowledge Resource”. In: *Frontiers in Artificial Intelligence and Applications*, IOS Press, vol. 178, 2008, pp.751-752.

[SCH07] Schönhofen, P.; Benczur, A.; Biro, I.; Csalogany, K. “Performing Cross-Language Retrieval with Wikipedia”. In: Cross-Language Evaluation Forum 2007 (CLEF 2007). Budapeste, Hungria, setembro 2007. Anais eletrônicos. Disponível em: http://www.clef-campaign.org/2007/working_notes/schonhofenCLEF2007.pdf. Acesso em: janeiro 2010.

[SMI01] Smith, B.; Welty, C. “FOIS introduction: Ontology---towards a new synthesis”. In: International Conference on Formal Ontology in Information Systems (FOIS01), 2001, pp.3-9.

[SOU04] Souza, R.; Alvarenga, L. “A Web Semântica e suas contribuições para a ciência da informação”**.** *Ciência da Informação,* vol. 33-1, junho 2004.

[SPI08] Spinellis, D.; Louridas, P. “The collaborative organization of knowledge”. *Communications of the ACM*, vol. 51-8, agosto 2008, pp.68-73.

[STR06] Strube, M.; Ponzetto, S. P. “WikiRelate! Computing Semantic Relatedness Using Wikipedia”. In: 21st National Conference on Artificial Intelligence (AAAI-06), 2006, pp.1419-1424.

[SUC08] Suchanek, F. M.; Kasneci, G.; Weikum, G. “YAGO: A Large Ontology from Wikipedia and WordNet”. *Web Semantics*, vol. 6-3, setembro 2008, pp.203-217.

[SYE08] Syed Z.; Finin T.; Joshi A. “Wikipedia as an Ontology for Describing Documents”. In: The International Conference on Weblogs and Social Media

(ICWSM08), 2008. Anais eletrônicos. Disponível em: http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-024.pdf. Acesso em: janeiro 2010.

[XAV09a] Xavier, C. C.; Lima, V. L. S. “Construção de uma Estrutura Ontológica de Domínio a partir da Wikipédia”. In: 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), 2009.

[XAV09b] Xavier, C. C.; Lima, V. L. S. “Extração de Estruturas Ontológicas a partir da Wikipédia em Língua Portuguesa”. In: 9º Workshop de Teses e Dissertações do Webmedia (WTD 2009), 2009. CD-ROM.

[ZES07] Zesch, T.; Gurevych, I. "Analysis of the wikipedia category graph for nlp applications”. In: 2nd TextGraphs Workshop (NAACL-HLT), 2007, pp.1-8.

[ZIR08] Zirn, C.; Nastase, V.; Strube, M. “Distinguishing between Instances and Classes in the Wikipedia Taxonomy”. In: 5th European Semantic Web Conference (ESCW08), 2008, pp.376-387

[WÖL06] Völkel, M.; Krötzsch, M.; Vrandecic, D.; Haller, H.; Studer, R. “Semantic Wikipédia”. In: 15th International Conference on World Wide Web (WWW2006), 2006, pp.585-594.

[WU07] Wu, F.; Weld, D. S. “Autonomously semantifying Wikipedia”. In: 16th ACM Conference on Conference on information and Knowledge Management (CIKM 2007), 2007, pp.41-50.

APÊNDICE A - LISTA DAS NÃO-CONFORMIDADES NO MAPEAMENTO DAS RELAÇÕES *LOCATED-IN* E *IS-A*

Relações *is-a* mapeadas na estrutura ontológica de Referência e não mapeadas da estrutura avaliada

[sitios_arqueologicos] *is-a* [atracoes_turisticas]
[miradouros] *is-a* [atracoes_turisticas]
[rotas_dos_vinhos] *is-a* [atracoes_turisticas]
[praias] *is-a* [atracoes_turisticas]
[marinas] *is-a* [atracoes_turisticas]
[monumentos] *is-a* [atracoes_turisticas]
[areas_verdes] *is-a* [atracoes_turisticas]
[cidades_da_antiguidade] *is-a* [atracoes_turisticas]
[jardins_zoologicos_por_pais] *is-a* [atracoes_turisticas]
[patrimonio_edificado] *is-a* [patrimonio_cultural_por_pais]
[bares] *is-a* [restaurantes]
[carnaval] *is-a* [turismo_cultural]
[regioes_de_turismo] *is-a* [turismo_por_pais]

Relações *is-a* mapeadas na estrutura ontológica avaliada e não mapeadas da estrutura de referência

[parques_nacionais] *is-a* [turismo_na_costa_rica]
[estacoes_de_esqui] *is-a* [resort]
[estacoes_de_esqui] *is-a* [turismo_na_america_do_sul]
[estacoes_de_esqui] *is-a* [turismo_por_pais]
[empresas_aereas] *is-a* [turismo_na_america_do_sul]
[empresas_aereas] *is-a* [turismo_por_pais]
[aeroportos] *is-a* [turismo_na_america_do_sul]
[aeroportos] *is-a* [turismo_por_pais]
[meios_de_hospedagem] *is-a* [empresas_do_setor_de_turismo]
[pousadas] *is-a* [turismo_em_portugal]
[hoteis] *is-a* [turismo_na_america_do_sul]
[hoteis] *is-a* [turismo_por_pais]
[turismo_rural] *is-a* [modelos_de_turismo]

[turismo_de_aventura] *is-a* [modelos_de_turismo]

[turismo_espacial] *is-a* [modelos_de_turismo]

[ecoturismo] *is-a* [modelos_de_turismo]

[sitios_arqueologicos] *is-a* [patrimonio_cultural]

[patrimonio_mundial_da_unesco] *is-a* [turismo_na_america_do_sul]

[patrimonio_mundial_da_unesco] *is-a* [turismo_por_pais]

[monumentos] *is-a* [patrimonio_cultural]

[monumentos] *is-a* [turismo_por_pais]

[cidades_da_antiguidade] *is-a* [sitios_arqueologicos]

[centros_de_convencoes] *is-a* [turismo_na_america_do_sul]

[centros_de_convencoes] *is-a* [turismo_por_pais]

[miradouros] *is-a* [turismo_em_portugal]

[regioes_de_turismo] *is-a* [turismo_em_portugal]

[patrimonio_edificado] *is-a* [turismo_em_portugal]

[patrimonio_edificado] *is-a* [turismo_na_america_do_sul]

[patrimonio_edificado] *is-a* [turismo_por_pais]

[praias] *is-a* [turismo_em_portugal]

[praias] *is-a* [turismo_por_pais]

[conservacao] *is-a* [turismo_por_pais]

[carnaval] *is-a* [turismo_na_america_do_sul]

[carnaval] *is-a* [turismo_por_pais]

[areas_verdes] *is-a* [turismo_na_america_do_sul]

[areas_verdes] *is-a* [turismo_por_pais]

[bares] *is-a* [turismo_por_pais]

[marinas] *is-a* [turismo_por_pais]

[jardins_zoologicos_por_pais] *is-a* [turismo_por_pais]

[rotas_dos_vinhos] *is-a* [turismo_por_pais]

Relações *located-in* mapeadas na estrutura ontológica de Referência e não mapeadas da estrutura avaliada

[turismo_por_pais] *located-in* [africa]

[turismo_por_pais] *located-in* [inglaterra]

[turismo_por_pais] *located-in* [irlanda_do_norte]

[patrimonio_mundial_da_unesco] *located-in* [africa]

[patrimonio_mundial_da_unesco] *located-in* [amazonia]

[patrimonio_mundial_da_unesco] *located-in* [nubia]
[reservas_da_biosfera] *located-in* [america_latina]
[atracoes_turisticas_por_pais] *located-in* [escocia]
[conservacao] *located-in* [argentina]
[parques_estaduais] *located-in* [liberia]
[lugares_historicos_registrados] *located-in* [columbia]
[sitios_arqueologicos] *located-in* [mesoamerica]
[sitios_arqueologicos] *located-in* [petra]
[transportes_publicos] *located-in* [israel]

Relações *located-in* mapeadas na estrutura ontológica avaliada e não mapeadas da estrutura de referência

[transportes_por_pais] *located-in* [grecia]
[parques_nacionais] *located-in* [bulgaria]
[parques_nacionais] *located-in* [liberia]
[transportes_publicos] *located-in* [supervia]

ANEXO A – RELATÓRIO SOBRE A CONSTRUÇÃO DO MODELO DE REFERÊNCIA ELABORADO PELA LINGUISTA SUSANA DE AZEREDO

OBSERVAÇÃO DE UMA EXTRAÇÃO ONTOLÓGICA DA ÁREA DE TURISMO A PARTIR DA WIKIPÉDIA: UM OLHAR LINGUÍSTICO

1. Introdução

O objetivo desse relatório é descrever as etapas envolvidas em uma observação, do ponto de vista linguístico, de uma proposta de extração de uma estrutura ontológica da área do Turismo. Essa extração partiu de uma estrutura de categorias da área de Turismo da Wikipédia de janeiro de 2009.

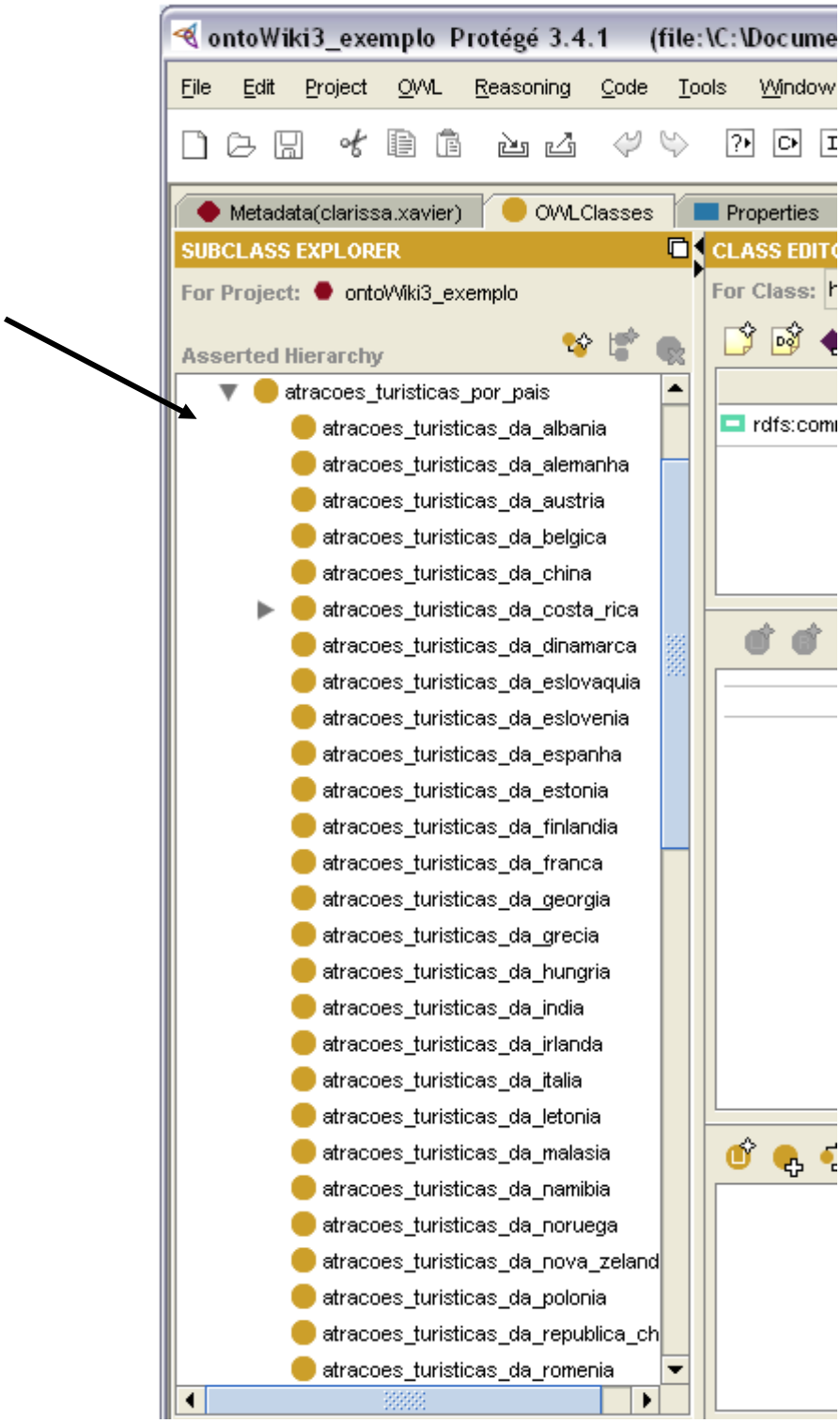
Em um primeiro momento, foi feita uma observação da instância *located in* existente em algumas classes e subclasses da estrutura. A partir dessa observação, passou-se para um segundo momento do trabalho, em que foi feita uma observação da organização semântica das classes e subclasses presentes.

A seguir, são listados os passos seguidos durante a observação, bem como as justificativas de algumas das escolhas feitas durante a organização semântica das classes e subclasses.

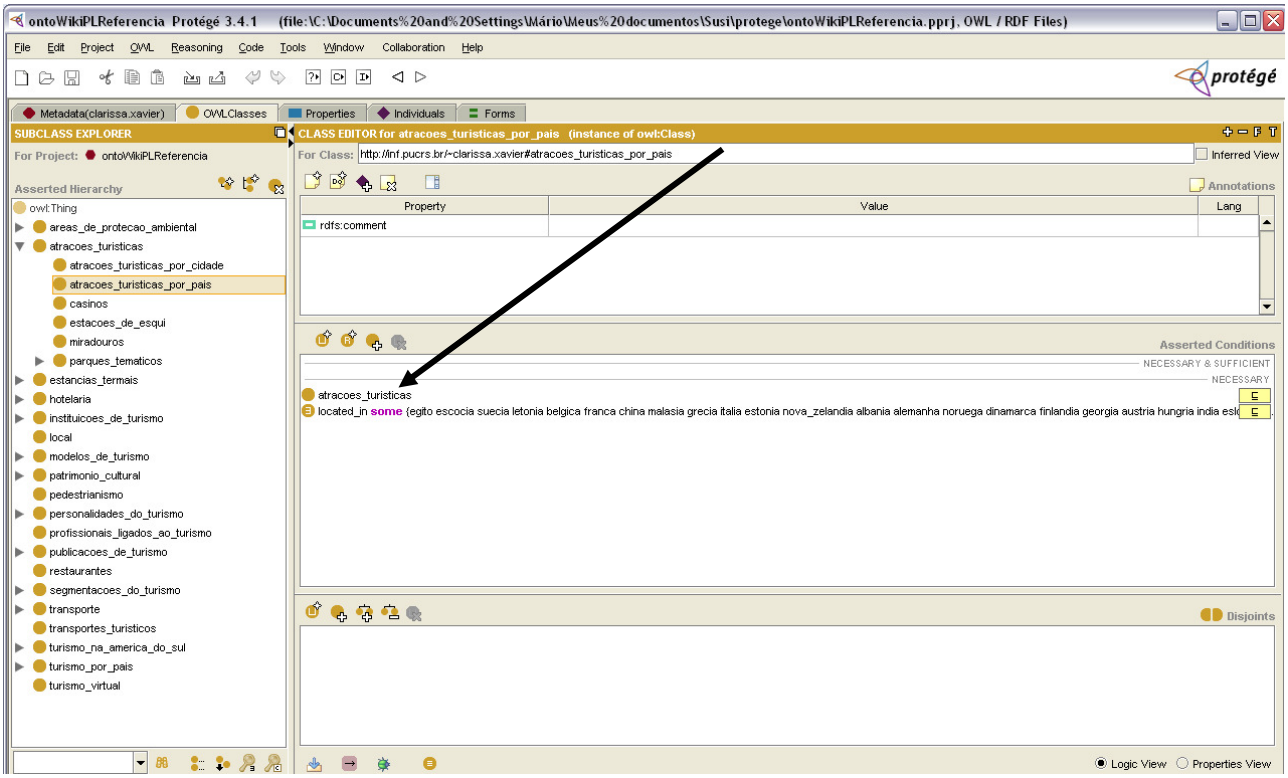
2. Observação da instância *located in*: um primeiro momento da observação

Em um primeiro momento da observação, foi solicitado que se observasse a instância *located in* presente em algumas classes e subclasses da estrutura ontológica extraída. Essa observação foi feita através da comparação de dois arquivos *.owl*.

Um dos arquivos apresentava as classes e subclasses conforme apareciam na lista de categorias da Wikipédia, sem a instância de local *located in.* Tem-se aqui diversas subclasses, conforme figura abaixo.



Em outro arquivo, as subclasses que representavam uma instância de lugar estavam representadas pela instância *located in,* conforme figura abaixo.



A comparação entre os dois arquivos permitiu verificar se o que era subclasse no primeiro arquivo estava, de fato, representado por uma instância *located in* no segundo arquivo, e se não havia algum outro tipo de instância estabelecida entre os dados apresentados. A observação foi feita classe por classe, subclasse por subclasse. Verificou-se que todas as instâncias de local estavam adequadas. Além disso, a autora do trabalho, Clarissa Castellã Xavier, optou por fazer algumas modificações entre as classes; tais modificações pareceram adequadas. Por exemplo, a classe **Estâncias Termais** possuía uma subclasse, **Termas**, a qual possuía uma subclasse **Termas de Portugal**. A autora apagou a subclasse **Termas de Portugal** e a colocou com uma instância *located in* na subclasse **Termas**. Essa mudança foi considerada adequada uma vez que segue uma relação de lugar e segue o critério utilizado em outras classes que apresentam instância de lugar.

Durante a observação da instância *located in,* foi-se observando a organização das classes e subclasses presentes na estrutura e percebeu-se que a maneira como estavam organizadas parecia muito fragmentada e aleatória. Foi proposto, então, a possibilidade de reorganizar as classes e subclasses da estrutura de uma forma que as relações semânticas entre elas parecessem menos fragmentadas. Assim, passou-se para um segundo momento da observação da estrutura ontológica. A seguir, relatamos como esse segundo momento se organizou.

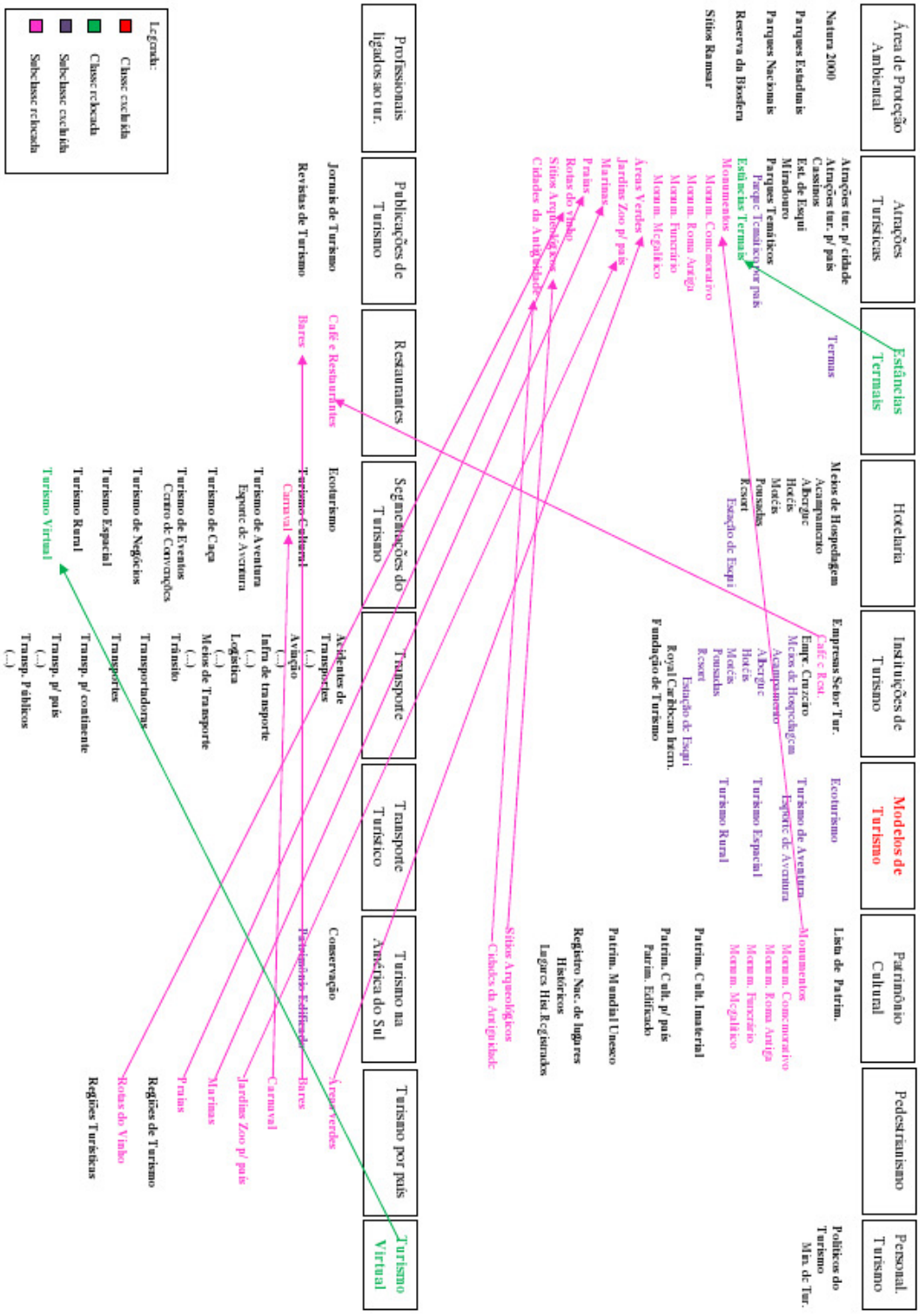
3. Observação da relação entre classes e subclasses: um segundo momento da observação

A observação das instâncias de local permitiu a familiarização com as classes e subclasses da

estrutura. A partir dessa primeira observação, algumas questões se levantaram. Por exemplo, por que havia uma classe chamada **Turismo Virtual** se havia uma classe chamada **Modelos de Turismo** com subclasses como **Turismo Rural, Turismo Espacial, Turismo de Aventura**, entre outras? Será, então, que a classe **Turismo Virtual** não deveria aparecer como uma subclasse de **Modelos de Turismo**? Além disso, por que havia uma classe chamada **Modelos de Turismo** e outra chamada **Segmentos do Turismo**, sendo que suas subclasses se repetiam? Não deveriam essas duas classes serem apenas uma? Questões como essas e outras permitiram pensar em uma reorganização das classes e subclasses da estrutura, a fim de que o tema *Turismo* ficasse mais bem representado naquelas categorias.

A partir daí, começou-se uma tentativa de reorganização da estrutura presente. À medida que se avançava nesta reorganização, novas questões surgiam. Sabe-se que as escolhas feitas são subjetivas e que podem ser questionadas, levando a uma reorganização mais adequada. No entanto, nesse momento, essa reorganização é uma tentativa de tornar a estrutura original menos fragmentada.

Abaixo, elaborou-se uma figura que mostra as alterações feitas na estrutura original.



Na figura acima, os quadrados representam as classes, e o que está logo abaixo de cada quadrado são as subclasses. O que está escrito em vermelho aponta para a exclusão de uma classe; o que está escrito em verde indica um deslocamento de uma classe para uma subclasse. As flechas apontam de onde as classes saíram e para onde foram colocadas. O que está escrito em rosa indica

uma subclasse que foi relocada e o que está em lilás indica uma subclasse que foi apagada. Da mesma forma, as flechas apontam de onde as subclasses saíram e para onde foram. As subclasses só eram excluídas quando apareciam mais de uma vez na estrutura. Por exemplo, a subclasse **Meios de Hospedagem** aparecia tanto na classe **Hotelaria** quanto na classe **Instituições de Turismo.** Nesses casos, uma delas foi apagada, visto que aparecendo mais de uma vez na estrutura, tornava-se redundante.

A reorganização das classes não foi aleatória. Para cada modificação há uma justificativa.

3.1 Alterações de classes e de subclasses: justificativas

3.1.1 A classe *Área de Proteção Ambiental*

A classe **Área de Proteção Ambiental** foi mantida como estava. Não foram feitas alterações nas suas subclasses.

3.1.2 A classe *Atrações Turísticas*

A classe **Atrações Turísticas** teve algumas alterações. Originalmente, essa classe tinha 6 subclasses (atrações turísticas por cidade, atrações turísticas por país, cassinos, estações de esqui, miradouros e parques temáticos). Foram incluídas nessa classe mais 9 subclasses (estâncias termais, monumentos, áreas verdes, jardins zoológicos por país, marinas, praias, rotas do vinhos, sítios arqueológicos e cidades da antiguidade), as quais estavam distribuídas em outras classes.

Estâncias Termais era uma classe que tinha como subclasse **Termas.** Pareceu adequado colocar **Estâncias Termais** como uma subclasse de **Atrações Turísticas** uma vez que parecem ser, conceitualmente, atrações turísticas. Não parece haver uma razão específica para haver uma classe somente para dar conta desse tema. Se assim fosse, não deveria existir classes para sítios arqueológicos, miradouros, praias, entre outros? Sendo assim, preferiu-se colocar **Estâncias Termais** como uma subclasse de **Atrações Turísticas.**

Manteve-se a instância *located in* existente nessa classe.

Monumentos era uma subclasse de **Patrimônio Cultural**. Relocamos essa subclasse para **Atrações Turísticas**, uma vez que nem todo monumento pode ser um Patrimônio Cultural, mas monumentos, de forma geral, podem ser uma atração turística. Assim, parecia mais adequado que a subclasse **Monumentos** estivesse em **Atrações Turísticas**. Reconhece-se que, nesse caso, é preciso considerar o que se entende por Patrimônio Cultural na lista de categorias da Wikipédia. No entanto, como essa informação não faz parte dos nossos dados, nos limitamos a entender a subclasse **Monumentos** de uma forma mais ampla.

Áreas verdes, **Jardins Zoológicos por País**, **Marinas**, **Praias** e **Rotas do Vinho** eram subclasses de **Turismo por País**. No entanto, quando se olha para a classe **Turismo por País,**

espera-se que ela apresente uma lista de países com suas possíveis atrações[28]. O que acontece é que, nessa classe, há uma lista de atrações turísticas que não estão organizadas por país. Assim, relocamos essas 5 subclasses para a classe **Atrações Turística**, a qual apresenta atrações turísticas de uma forma geral, sem organizá-las por país.

Sítios Arqueológicos era uma subclasse de **Patrimônio Cultural** e **Cidades da Antiguidade** era uma subclasse de **Sítios Arqueológicos**. Deslocamos **Sítios Arqueológicos** e **Cidades da Antiguidade** para subclasse de **Atrações Turísticas**. No caso de **Sítios Arqueológicos**, essa mudança justifica-se, uma vez que nem todos os sítios arqueológicos podem ser patrimônio cultural, mas podem ser atrações turísticas. Além disso, **Cidades da Antiguidade** como subclasse de **Sítios Arqueológicos** dava a entender que todas as cidades da antiguidade eram sítios arqueológicos, o que pode não ser real. Assim, ao invés de ser uma subclasse de **Sítios Arqueológicos**, colocamos **Cidades da Antiguidade** como uma subclasse de **Atrações Turísticas,** pois muitas cidades da antiguidade são atrações turísticas.

3.1.3 A classe *Estâncias Termais*

Essa classe foi colocada como subclasse de **Atrações Turísticas.** A subclasse **Termas** foi excluída e a instância *located in* foi mantida. Não parecia haver necessidade de uma classe específica para tratar de estâncias termais. Assim, visto que havia uma classe chamada **Atrações Turísticas** e visto que estâncias termais são atrações turísticas, colocamos a classe **Estâncias Termais** como subclasse de **Atrações Turísticas.**

3.1.4 A classe *Hotelaria*

A classe **Hotelaria** teve uma modificação que foi o apagamento da subclasse **Estação de Esqui.** Essa subclasse foi excluída uma vez que ela já estava presente como subclasse em **Atrações Turísticas,** tornando-se redundante.

3.1.5 A classe *Instituições de Turismo*

Nessa classe, retirou-se a subclasse **Meios de Hospedagem,** uma vez que ela já estava presente em Hotelaria. Além disso, a subclasse **Café e Restaurantes** foi relocada para a classe **Restaurantes** por dois motivos. O primeiro é que cafés e restaurantes não são *instituições de turismo* e parecia haver uma diferença semântica grande entre a subclasse **Cafés e Restaurantes** e as outras subclasses. Cafés e restaurantes podem ser considerados locais que movimentam o turismo ou que contribuem para o turismo, mas não *instituições*. Essa palavra, *instituições*, parece

[28] Talvez fosse possível ter uma classe chamada **Turismo por País** com uma subclasse chamada **Atrações Turísticas por País.**

dar conta de algo mais formalizado. O segundo motivo é que a classe **Restaurantes** estava vazia e essa categoria pode ser considerada um item importante para o setor de *Turismo*.

3.1.6 As classes *Modelos de Turismo* e *Segmentações de Turismo.*

Todas as subclasses da classe **Modelos de Turismo** foram excluídas porque estavam presentes na classe **Segmentos de Turismo.** Com isso, excluímos a classe **Modelos de Turismo.** Essa escolha de exclusão da classe foi única e singular. Os motivos da nossa escolha foram baseados em uma breve pesquisa no site do Ministério do Turismo e no Google. No site do Ministério do Turismo, há publicações que utilizam a expressão **Segmentos do Turismo** ou **Segmentações do Turismo** para se referir aos diferentes tipos de turismo como, por exemplo, Turismo Rural, Turismo de Negócios, Ecoturismo, etc, exatamente como aparece na nossa estrutura ontológica. Em uma breve pesquisa no Google, é possível perceber que essa expressão é utilizada com esse mesmo objetivo, também, em revistas de turismo.

Por outro lado, não encontramos a expressão **Modelos de Turismo** referindo-se aos diferentes tipos de turismo. Uma busca no Google revelou que a expressão **Modelos de Turismo** é muito ampla e faz referência até mesmo a alguns modelos de carro. Parece que, para a área de Turismo, a expressão *Segmentações de Turismo* é a terminologia utilizada para se referir às ramificações da área. Por essa razão optamos por excluir a classe **Modelos de Turismo** e deixar a classe **Segmentações de Turismo** com as subclasses que fazem referência a diferentes tipos de turismo.

Além disso, foi incluída como subclasse, na classe **Segmentações de Turismo,** a classe **Turismo Virtual**, uma vez que pareceu ser uma das ramificações da área do turismo. Com isso, a classe **Turismo Virtual** deixou de existir.

3.1.7 A classe *Patrimônio Cultural*

Nessa classe, foram retiradas as subclasses **Monumentos e Sítios Arqueológicos.** Essas subclasses foram colocadas em **Atrações Turísticas.** Essa escolha foi movida pelo fato de que nem sempre os monumentos e os sítios arqueológicos podem ser considerados patrimônios culturais, embora possa ser considerados como atrações turísticas. Como não se sabe qual o conceito de Patrimônio Cultural adotado pela Wikipédia, resolveu-se entender os monumentos e os sítios arqueológicos de uma forma mais ampla e colocá-los em **Atrações Turísticas**.

3.1.8 A classe *Pedestrianismo*

Essa classe estava vazia. Não foi feita nenhuma modificação nessa classe. Em um primeiro momento, pensou-se em excluí-la. No entanto, o fato de estar vazia não significa que não

seja uma categoria importante para a área do turismo e que não possa ser preenchida mais tarde.

3.1.9 A classe *Personalidades do Turismo*

Nessa classe não foi feita nenhuma alteração.

3.1.10 A classe *Profissionais Ligados ao Turismo*

Não foi feita nenhuma alteração nessa classe. Embora vazia, não foi apagada uma vez que é uma classe importante e pode vir a ser preenchida.

3.1.11 A classe *Publicações de Turismo*

Essa classe não foi modificada

3.1.12 A classe *Restaurantes*

A classe **Restaurantes** estava vazia. Em um primeiro momento, pensou-se em apagar essa classe e colocá-la como subclasse em **Empresas do Setor do Turismo** (subclasse de Instituições de Turismo). No entanto, estavam espalhadas, em outras classes, subclasses que estavam relacionadas com alimentação. Assim, resolvemos colocar todas essas subclasses juntas sob a classe **Restaurantes,** pois esse parece ser um item relevante para a área de Turismo. Assim, trouxemos para a classe **Restaurantes** a subclasse **Café e Restaurantes** (subclasse de **Empresas do Setor do Turismo**) e a subclasse **Bares** (subclasse de **Turismo por país**).

No entanto, entendemos que talvez o nome colocado para essa classe (**Restaurantes**) não seja o mais adequado para dar conta de tudo o que ela abrange ou venha a abranger como, por exemplo, cafés e bares. Alguém talvez questione que cafés e bares não são restaurantes. É verdade. Mas, na divisão que temos, parece ser importante uma categoria que se refira a restaurantes, bares, cafés e ainda outros setores que tratem de alimentação.

3.1.13 A classe *Segmentações de Turismo*

Considerações sobre as alterações feitas nessa classe foram consideradas no item 3.1.6.

3.1.14 As classes *Transporte* e *Transportes Turísticos*

Não foram feitas alterações nessas classes. Foi pensado na exclusão de uma das classes, pois, em um primeiro momento, pareceu ter duas classes que representavam a mesma coisa. No entanto, em um segundo momento, pareceu que as classes parecem representar segmentos diferentes do Transporte. Enquanto a classe **Transportes** parece representar a categoria *transportes*

de uma forma mais ampla, incluindo subclasses como trânsito, infraestrutura do transporte, meios de transporte, etc, a classe **Transportes Turísticos,** embora vazia, pode se referir a tipos de transportes mais específicos como, por exemplo, o balão, o Tuk Tuk na India, o moto taxi no Peru, entre outros. Assim, embora a classe **Transportes Turística** esteja vazia, optou-se por deixá-la ali para que possa ser preenchida.

3.1.15 A classe *Turismo na América do Sul*

Nessa classe, foi excluída a subclasse **Patrimônio Edificado**, uma vez que já estava presente como subclasse em **Patrimônio Cultural –> Patrimônio Cultural por país –> Patrimônio Edificado.**

3.1.16 A classe *Turismo por país*

Essa classe foi bastante alterada. Sob essa classe, espera-se, talvez, uma relação de países com suas especificações (ver item 3.1.2). No entanto, as subclasses de **Turismo por país** parecem estar mais relacionadas com atrações turísticas. Assim, a grande maioria das subclasses foi relocada para a classe **Atrações Turísticas.** São elas: **áreas verdes, praias, marinas, jardins zoológicos por país** e **rotas do vinho.**

Além disso, a subclasse **Bares** foi colocada como subclasse da classe **Restaurantes,** e a subclasse **Carnaval** foi colocada como subclasse de **Segmentações do Turismo –> Turismo Cultural -> Carnaval.**

3.1.17 A classe *Turismo Virtual*

Essa classe foi colocada como subclasse em **Segmentações do Turismo.**

4. As subclasses *Atrações turísticas por cidade* e *Atrações turísticas por país*: uma consideração

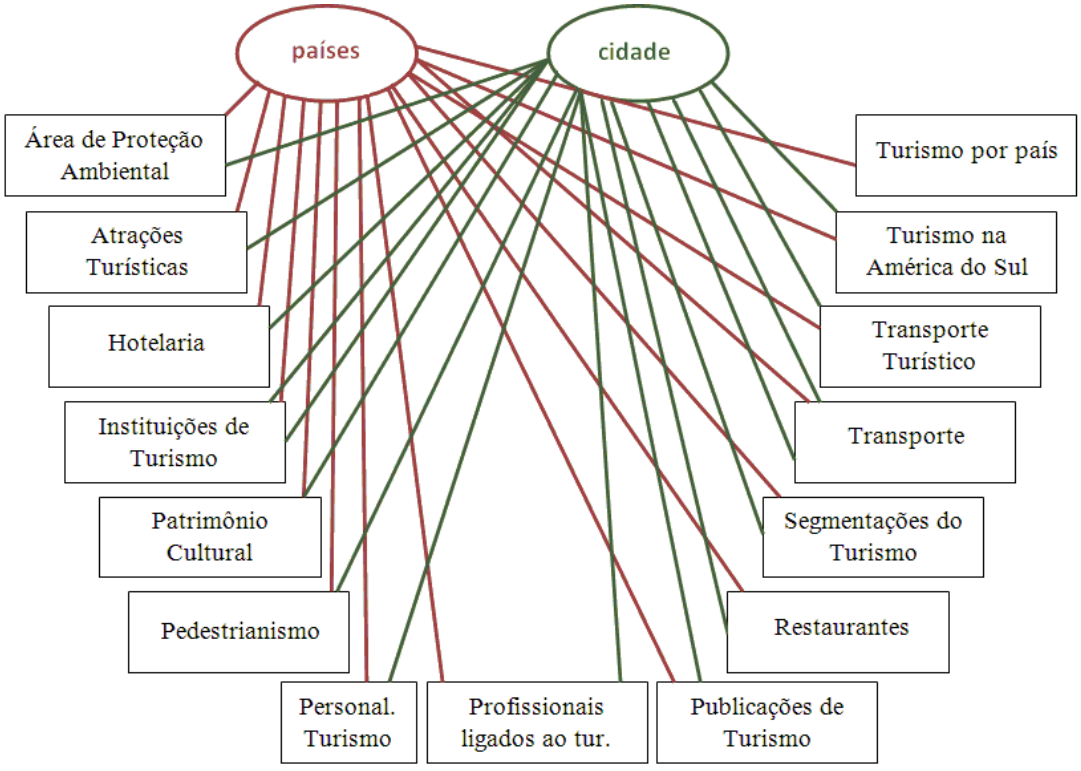
Na classe **Atrações Turísticas**, tem-se como subclasse **Atrações Turísticas por país** e **Atrações Turísticas por cidade.** Embora essas subclasses tenham sido mantidas no seu lugar, gostaríamos de fazer algumas considerações sobre elas.

Pareceu que essas duas subclasses diferem muito das outras subclasses encontradas em **Atrações Turísticas.** Parece que essas duas subclasses não fazem parte dessa classe; parecem ser algo maior do que **Atrações Turísticas.** No entanto, parece também estranho colocar essas subclasses como classes na nossa lista. Assim, sem saber exatamente onde colocá-las, resolvemos deixá-las onde estão.

Porém, parece que as subclasses **Atrações Turísticas por país** e **Atrações Turísticas por**

cidade são algo maior do que a própria estrutura ontológica como um todo. Pensou-se em estabelecer essas subclasses como uma supercategoria, ou como uma nova raiz. Com isso, diferentes informações sobre diferentes países e cidades poderiam ser acionadas através da escolha de um país ou cidade. Isso resultaria em um cruzamento de informações. Talvez, com isso os nomes **Atrações Turísticas por País** e **Atrações Turísticas por cidade** não sejam os melhores. Talvez se poderia ter apenas **Países, Cidades**, entre outros. Tentando representar essa idéia, surgiu a palavra *matriz ontológica*, cunhada pela aluna Clarissa C. Xavier. Essa expressão tenta traduzir a idéia de um cruzamento de informações, tendo como supercategoria uma lista de países, de cidades, entre outros.

Tentando representar graficamente essa idéia, temos a figura abaixo:



A figura acima tenta traduzir a sugestão de retirar as subclasses **Atrações Turísticas por País** e **Atrações Turísticas por cidade** e colocá-las como algo superior a estrutura. A idéia seria fazer com que por meio da seleção de um país, uma cidade, um estado, uma região, um continente, etc., possa-se acessar diferentes informações sobre o local escolhido.

5. Considerações finais

Ao final desta observação, muitas questões ficaram ainda sem respostas. Por exemplo, o que se entende por **Patrimônio Cultural** na lista de categorias da Wikipédia? Ou, então, o que se entende por cidades da antiguidade? Além disso, qual a diferença entre a classe **Turismo**

por país e a subclasse **Atrações turísticas por país**? Essas questões vão além do trabalho proposto pela aluna Clarissa C. Xavier, mas são questões pertinentes para a proposta de uma futura ontologia da área de Turismo.

As respostas de questões como as acima citadas podem fazer com que o lugar de algumas classes e subclasses seja repensado. Isso revela o quão subjetiva é a observação feita. Ela foi feita a partir da visão de um profissional da Terminologia que percebeu uma determinada relação entre as classes e, a partir dessa relação, distribuiu as classes e subclasses conforme essa visão. É possível que outro profissional da Terminologia perceba essas relações de outra forma e, considerando outros fatores, sugira outras escolhas.

Essas diferenças revelam também que a compilação manual é complexa. Surge, então, outra questão: se uma compilação manual é complexa, o que se esperaria de uma compilação automática? Considerando que nem sempre o automático reproduz exatamente o natural, mas o imita de outras formas, percebe-se que diferentes caminhos podem ser criados para se tentar imitar o natural de uma forma automática.

Para além dessas questões, salienta-se o uso do Protégé. O software foi bastante útil para a reorganização da estrutura uma vez que permite o deslocamento das classes ou subclasses de uma forma bem simples. É preciso estar atento para que, nesses deslocamentos, não se apague nenhuma relação existente entre as classes. No entanto, o Protégé foi útil somente na parte final da observação quando já se sabia quais as alterações que seriam feitas. Para se ter uma visão geral da estrutura e para se pensar na reordenação das classes, foi necessário sair do Protégé e elaborar manualmente o desenho na página 5. Com aquele desenho era possível mudar as classes de lugar, voltar atrás em algumas decisões, apagar algumas mudanças e repensá-las de uma forma mais rápida. Durante a observação era muito comum voltar atrás em algumas decisões. Por exemplo, em um dado momento, achou-se adequado apagar uma classe. No entanto, um pouco depois se pensou que esse apagamento não seria adequado e se voltou atrás na decisão. Com o desenho, simplesmente escrevíamos a classe novamente no papel. Com o Protégé, isso não era possível, pois mudanças como os apagamentos eram definitivas. Para refazer essas mudanças, era necessário começar novamente uma nova estrutura e fazer todas as alterações. Assim, optou-se por fazer um desenho que representasse a estrutura existente no Protégé e no qual se pudesse apagar, reescrever e mudar. Somente após terem sido feitas todas as alterações, passou-se a nova estrutura para o Protégé. Nesse momento, o Protégé foi muito prático, pois permite que as alterações sejam feitas de uma forma bem simples.

Por fim, levanta-se a questão: como seria a elaboração de uma estrutura ontológica de turismo para outras línguas? Seria possível utilizar a Wikipédia em outras línguas para isso? Uma breve observada na Wikipédia em Inglês e em Francês revela algumas diferenças entre as categorias

presentes em língua portuguesa, em língua inglesa e em língua francesa. Talvez em outras línguas as categorias presentes em nossa estrutura ontológica estejam organizadas de uma forma diferente. Uma comparação com estruturas ontológicas do turismo em outras línguas pode ajudar a responder algumas das perguntas levantadas ao longo desta observação.

Um trabalho de comparação com outras estruturas ontológicas, que envolveram escolhas de outras pessoas, ajuda a avaliar as nossas decisões e a aprimorar, assim, a nossa estrutura.