UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Anotação semântica baseada em ontologia para análise de entrevistas dos atletas olímpicos brasileiros

Rovilson de Freitas

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Rovilson de Freitas

Anotação semântica baseada em ontologia para análise de entrevistas dos atletas olímpicos brasileiros

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Elaine Parros Machado de Sousa

USP – São Carlos
Novembro de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

DE FREITAS, ROVILSON
D278a Anotação semântica baseada em ontologia para análise de entrevistas dos atletas olímpicos brasileiros / ROVILSON DE FREITAS; orientadora Elaine Parros Machado de Sousa. -- São Carlos, 2022.
102 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Ontologia de Domínio. 2. Anotação Semântica. 3. Mineração de Textos. I. Parros Machado de Sousa, Elaine, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rovilson de Freitas

Ontology-based semantic annotation for analysis of interviews with brazilian olympic athletes

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Elaine Parros Machado de Sousa

USP – São Carlos
November 2022

Este trabalho é dedicado à todos os pesquisadores brasileiros,
que se superam diariamente.
Ciência salva vidas.

AGRADECIMENTOS

Agradecimentos a minha orientadora Prof[a] Dr[a]. Elaine Parros Machado de Sousa, pela infinita paciência na condução desse trabalho. Sua sensibilidade em compreender os imprevistos impostos ao longo do caminho trouxeram segurança e conforto. Sem seu olhar cuidadoso, muitos detalhes teriam passado despercebidos. Minha eterna gratidão por todo auxílio dispensado.

Também agradecimentos muito especiais à Prof[a] Dr[a]. Katia Rubio, por ter me recebido no Grupo de Estudos Olímpicos da Universidade de São Paulo (GEO/USP) em 2016 e ter confiado a mim os dados tão preciosos ao Grupo. Conselhos não apenas para a conclusão desse trabalho, mas que serão levados para toda a vida. Não poderia deixar de citar todos os membros do GEO. Esse trabalho só existe graças ao árduo trabalho desenvolvido por vocês ao longo de muitos anos.

À Professora Solange Oliveira Rezende, que abriu as portas do ICMC junto com minha orientadora e o colega Ruan Neves, para que esse projeto fosse possível. Menção especial aos professores Ricardo Marcondes Marcacini e Roberta Akemi Sinoara, sempre dispostos a nos auxiliar no que fosse preciso.

À todos os meus familiares, que sempre me incentivaram pelo caminho da educação. Aos amigos pelo apoio e pela compreensão quando a ausência foi necessária.

À CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro.

Ao Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS), por permitir que me licenciasse sem prejuízos ao salário para o desenvolvimento desse trabalho.

Aos atletas olímpicos brasileiros, que sempre foram uma fonte de inspiração. Esse trabalho é mais uma contribuição para que a história de vocês jamais seja esquecida.

“Achei a palavra atleta bonita e decidi que queria ser um”
(Adhemar Ferreira da Silva, Bicampeão Olímpico - 1952/1956)

RESUMO

FREITAS, R. **Anotação semântica baseada em ontologia para análise de entrevistas dos atletas olímpicos brasileiros**. 2022. 102 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Normalmente, pesquisas acadêmicas coletam um grande acervo de dados. Esses dados, ao longo do tempo, precisam ser acessados e manipulados pelos pesquisadores, de acordo com a natureza de sua investigação. É fundamental que esses dados estejam disponibilizados de maneira simples, com algum suporte computacional para facilitar o trabalho dos pesquisadores. A realidade da pesquisa, de maneira geral, corresponde a recursos escassos e, portanto, o tempo precisa ser otimizado. O presente trabalho propõe uma possível solução que apoie tarefas de análise e descoberta de conhecimento a partir do acervo do Grupo de Estudos Olímpicos da Universidade de São Paulo, utilizando estratégias de anotação semântica baseada em ontologia, aliada com técnicas de mineração de texto. Para isso, foi desenvolvida uma ontologia de domínio chamada OntOlympic, que serviu de base para o processo de anotação semântica. As entrevistas passaram por um processo de mineração de textos (agrupamentos), com e sem anotação semântica. Os resultados mostram que os grupos formados a partir das entrevistas anotadas tem uma tendência de serem melhores agrupamentos do que os grupos formados pelas entrevistas não anotadas. Os resultados, tanto do índice de avaliação (índice de Davies-Bouldin), quanto da análise dos grupos formados se demonstraram ligeiramente melhores. Como perspectiva futura, outros grupos que trabalham com a mesma dinâmica podem utilizar os processos desse trabalho, além de abrir perspectiva de outros testes na área de mineração de textos.

Palavras-chave: Ontologia de Domínio, Anotação Semântica, Mineração de Textos.

ABSTRACT

FREITAS, R. **Ontology-based semantic annotation for analysis of interviews with brazilian olympic athletes**. 2022. 102 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Typically, academic research collects a large body of data. This data, over time, needs to be accessed and manipulated by researchers, according to the nature of their investigation. It is critical that these simple data be available in a computer-supported manner to facilitate the work of researchers. The reality, general, research, scarce resources and therefore time needs the optimization to be. The work proposed by the University of São Paulo is a possible solution and supports the tasks of analysis and knowledge discovery from text mining techniques. For this, an Olympic domain ontology was developed, which served as the basis for the semantic annotation process. The interviews interviewed by a mining mining process (clusters), with and without ananotics. The results show that the groups that form the annotated interviews tend to be better groups than the groups that form the unannotated interviews. The results of both the evaluation index (Davies-Buldin index) and the formed groups compare the best of the analysis. As a future perspective, other test groups that work with the same can use the processes of this work, in addition to opening perspective of other text mining groups.

Keywords: Domain Ontology, Semantic Annotation, Text Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

GEO	Grupo de Estudos Olímpicos
ICMC	Instituto de Ciências Matemáticas e de Computação
ISI	Information Sciences Institute
Labic	Laboratório de Inteligência Artificial
PRAXIS	Processo para Aplicativos extensíveis e Interativos
RUP	Rational Unified Process
TOVE	Toronto Virtual Enterprise
UP	Unified Process
URIs	Uniform Resource Identifier
USP	Universidade de São Paulo

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

As pesquisas acadêmicas no Brasil, promovidas principalmente por Instituições e Universidades públicas, representam importante contribuição para o avanço da ciência. Nas mais diferentes áreas, cada uma com suas características próprias e utilizando das mais variadas ferramentas metodológicas, pesquisas buscam trazer benefícios para a comunidade (melhoria da qualidade de vida das pessoas, desenvolvimento de novos produtos, serviços, entre outros). Além disso, as pesquisas podem permitir que novas reflexões sejam feitas sobre temas já visitados, trazendo novos olhares, perspectivas e soluções.

De modo geral, uma das fases iniciais e mais importantes para a pesquisa é a coleta de dados. Em muitos casos, essa coleta é feita por meio de entrevistas. Marconi e Lakatos (1999) definem entrevista como o encontro entre duas pessoas, a fim de que uma delas obtenha informações a respeito de determinado assunto, mediante uma conversação de natureza profissional.

As entrevistas são realizadas com sujeitos que podem de alguma maneira contribuir para a pesquisa, trazendo elementos de sua vivência em alguma atividade específica. A experiência pessoal, em determinadas situações, é fundamental para o entendimento do todo sobre aquela atividade. Ninguém melhor para falar sobre algo do que aquele que viveu ou presenciou o fato. A entrevista pode ter diferentes abordagens: realizada de maneira padronizada, estruturada, sem padrão, não-estruturada ou por meio de painel. Tudo depende do objetivo. Em alguns casos, se opta por algum modelo, e em outros pela mescla deles.

Atualmente, com a melhora e barateamento de equipamentos, é possível filmar ou gravar as entrevistas de maneira mais simples em formato digital. Após a entrevista, ela pode ser transcrita para algum formato de texto e armazenada em disco. Essa transcrição precisa estar disponível de maneira simples e o mais organizada possível. No entanto, pesquisadores em diferentes áreas nem sempre têm conhecimento avançado em ferramentas tecnológicas específicas, e muitos limitam-se ao básico da utilização do computador como usuário. Gerenciar

e facilitar esse acesso é fundamental para que a pesquisa se desenvolva da melhor maneira possível. A tecnologia não deve ser um problema nesse processo, mas sim uma ferramenta de apoio ao pesquisador.

Além da disponibilidade, a tecnologia pode proporcionar outras ferramentas para lidar com essas entrevistas. A grande quantidade de texto disponível pode ser uma oportunidade de aplicar estratégias de mineração de textos, por exemplo. Com a mineração é possível descobrir padrões e similaridades entre os textos selecionados. Isso pode facilitar ainda mais a análise do pesquisador, servindo como atalho para seu objetivo final. Vale lembrar que o tempo é um recurso bastante precioso aos pesquisadores. Ao invés da leitura de muitas (ou todas as) entrevistas, trabalho muitas vezes desnecessário, uma ferramenta pode orientar o pesquisador para que sua leitura, análise e conclusões sejam em menor tempo possível e de maior qualidade.

1.1 Motivação e contextualização do problema

O Grupo de Estudos Olímpicos (GEO)[1] da Universidade de São Paulo (USP), liderado pela Prof.ª Dr.ª Katia Rubio[2], atua no estudo do olimpismo nas mais diferentes perspectivas: história do esporte, preservação da memória, narrativas biográficas, psicologia do esporte, educação olímpica, entre outros. Rubio (2015) afirma que o objetivo principal da pesquisa é entender o que faz pessoas chegarem ao limite de seus corpos e de sua emoção na busca de um movimento que pode eternizá-las, especialmente numa edição de Jogos Olímpicos. Não existe nada mais importante para a grande maioria dos atletas do que estar nessa competição em especial.

No GEO, uma das formas de coleta de dados é através de entrevistas. Porém, as entrevistas são realizadas de maneira não-estruturada. Esse método de abordagem na entrevista tem o nome de narrativas biográficas. Nessa modalidade, o fluxo de pensamento do entrevistado (em sua quase totalidade, atletas olímpicos) é parte importante do processo. Não existe um roteiro pré-determinado e parte-se de uma pergunta desencadeadora: *Me conte sua história de vida*. Os assuntos pertinentes surgem a partir da fala desse entrevistado. Sobre as narrativas biográficas:

> A busca inicial pelas histórias de vida deu-se pelo entendimento de que era preciso permitir que os atletas organizassem suas lembranças, trajetórias e memórias, de forma a relatar não apenas os componentes objetivos dessa vivência como as principais conquistas, as participações olímpicas, quem os influenciou, mas principalmente os componentes de ordem pessoal e subjetivos carregados de afetividade e emoções de toda espécie. A associação entre essas duas instâncias traria as pistas necessárias para o entendimento da complexidade de uma pessoa que alia a condição de um nível de habilidade motora extraordinária à condição humana ordinária, que partilha das mesmas angústias e expectativas dos demais que vivem em sociedade.(Rubio (2014))

[1] https://www.olimpianos.com.br/
[2] http://lattes.cnpq.br/0941910739814664

Para Zimmermann (2019), a história oral e as histórias de vida ajudam a criar uma imagem mais verdadeira do passado e da mudança do presente, documentando as vidas e os sentimentos e muitas outras imagens escondidas da história. Após a entrevista, ela é transcrita e salva em documentos do Microsoft Word [3]. Ao longo dos anos, são mais de mil entrevistas transcritas no acervo do grupo. Sobre o objetivo do acervo, Rosina (2018) diz que o projeto tem como objetivo catalogar a história olímpica brasileira por meio da história de vida dos atletas que protagonizaram esses momentos.

São aproximadamente vinte anos em que o Grupo busca as narrativas biográficas de todos os atletas olímpicos brasileiros. As entrevistas podem ser realizadas com os atletas, com seus familiares (no caso dos já falecidos) e/ou atletas contemporâneos de modalidade.

A primeira participação oficial dos atletas brasileiros em Jogos Olímpicos ocorreu em 1920. O acervo de entrevistas do GEO reúne, aproximadamente, mil entrevistas, dos mais de dois mil atletas que participaram de pelo menos uma edição olímpica. Cada entrevista tem, em média, 10 páginas transcritas em arquivos texto. A cada nova edição olímpica (realizada de quatro em quatro anos), em média, cem novos atletas são adicionados a essa lista. Além das entrevistas propriamente ditas, também são coletados dados objetivos (contatos, data e local de nascimento, informações sobre formação acadêmica e experiência profissional, entre outros), que podem ser utilizadas no futuro por pesquisadores.

O GEO aponta uma dificuldade na utilização efetiva dessas entrevistas, pois estão armazenados em documentos de texto do Microsoft Word. Não existe nenhuma organização estruturada ou lógica, seja por modalidade, por edição olímpica, por sexo ou qualquer outra categoria. Portanto, o acesso aos dados dos atletas ou às entrevistas é todo manual (busca e leitura). Ou seja, é necessário realizar as buscas entre as mais de mil entrevistas usando os recursos do sistema operacional (busca pelo nome do arquivo, por exemplo) ou de algum processador de textos (busca por palavra dentro de cada arquivo). Consultas mais complexas, que permitam, por exemplo, estabelecer relações entre fatos relatados por diversos atletas, tornam-se extremamente custosas, e em muitos casos, inviáveis.

Outro ponto importante é que tanto os dados quanto as entrevistas estão disponíveis para qualquer integrante do grupo de pesquisa que tenha acesso ao laboratório. Por se tratar de armazenamento de dados pessoais, e eventualmente, alguma entrevista com conteúdo restrito, essa vulnerabilidade pode causar problemas relacionados a confidencialidade, consistência, integridade e validade dos dados. São exemplos a perda de dados (por remoção ou alteração indevida) e a divulgação de informação confidencial.

As entrevistas, em particular, podem ser fonte de inúmeras descobertas posteriores. Segundo TAN (1999), uma das abordagens promissoras é a utilização de técnicas de Mineração de Textos para extração de conhecimento em grandes coleções textuais. Padrões e relações

[3] https://www.office.com/

entre os textos, descobertos de modo eficiente, podem apoiar trabalhos de pesquisa baseados no acervo de entrevistas. Em pesquisas do GEO, por exemplo, o perfil psicológico do atleta é analisado, a partir de sua fala na entrevista. Resultados iniciais, de um trabalho desenvolvido por pesquisadores do GEO em parceria com pesquisadores do Laboratório de Inteligência Artificial (Labic) do Instituto de Ciências Matemáticas e de Computação (ICMC), mostram o potencial da utilização de padrões extraídos dos textos em análises de personalidade RUBIO *et al.* (2019).

Portanto, soluções computações que deem suporte ao uso dessas entrevistas são muito importantes para o grupo de estudos. Essas soluções devem proporcionar aos pesquisadores um melhor uso de seu tempo. Importante destacar que, apesar do suporte computacional, os processos realizados pelo humano ainda será necessário em algum momento.

1.2 Objetivo do Trabalho

Diante do cenário atual do acervo de dados de atletas olímpicos do GEO, este trabalho tem por objetivo geral apresentar uma solução que apoie tarefas de análise e descoberta de conhecimento a partir desse acervo, utilizando estratégias de anotação semântica baseada em ontologia, aliada com técnicas de mineração de texto (*Text Mining*). O objetivo geral deste trabalho de mestrado pode ser detalhado em três objetivos mais específicos:

1. Propor e implementar uma Ontologia para o domínio de aplicação, visando suporte à recuperação de informação com um viés mais semântico. Pretende-se que essa Ontologia considere as especificidades do GEO, os temas mais trabalhados ao longo dos anos, e possíveis assuntos futuros que podem vir a ser pesquisados pelo grupo.

2. A partir dessa ontologia, utilizar a técnica de anotação semântica, com objetivo de enriquecer as entrevistas com os temas de interesse e relevantes para o GEO. Como hipótese, esse passo poderá permitir que as entrevistas com assuntos relacionados possam ser percebidas de maneira automatizada, minimizando o processo manual e repetitivo.

3. Aplicar técnicas de mineração de textos, em particular agrupamento (clustering). Existe a hipótese de que o procedimento de anotação semântica baseada na ontologia contribui para aprimorar o agrupamento das entrevistas.

1.3 Contribuições principais

O trabalho propõe uma ontologia chamada OntOlympic, baseada nas demandas do Grupo de Estudos Olímpicos da Universidade de São Paulo. Essa ontologia, serve de base para o processo de anotação semântica em entrevistas pertencentes ao acervo do grupo. As entrevistas foram realizadas com atletas olímpicos brasileiros. Com as entrevistas anotadas, houve a aplicação de técnicas de mineração de textos, especificamente agrupamentos. Esses agrupamentos

fornecem subsídios para auxiliar os pesquisadores do GEO, mostrando possíveis assuntos de pesquisa ou confirmando hipóteses já debatidas. Com isso, pretende proporcionar uma menor quantidade de trabalho manual, e menos tempo dispensado nas análises. Embora este trabalho tenha como motivação os problemas de um grupo de pesquisa em especial (GEO), as soluções resultantes deste mestrado poderão, potencialmente, beneficiar trabalhos de outros grupos em cenários similares. Outros grupos de pesquisa que trabalham com coleta de dados baseada em entrevistas não-estruturadas em formato de texto e que não possuem suporte computacional adequado, podem adaptar à realidade abordada nesse trabalho.

1.4 Organização do documento

Essa dissertação está organizada da seguinte maneira: no capítulo 2, são apresentados os principais conceitos desse trabalho (ontologia, anotação semântica e mineração de textos). No capítulo 3, os trabalhos relacionados. No capítulo 4, o passo a passo da criação da ontologia OntOlympic. No capítulo 5, uma explicação do processo de anotação semântica e mineração de texto para esse trabalho. No capítulo 6, os experimentos realizados, e no capítulo 7, as considerações finais. Por último, temos as referências bibliográficas e apêndices.

CAPÍTULO

2

CONCEITOS

Nesse capítulo, temos as principais definições do conceito de ontologia, no seu aspecto amplo e especificamente para a área de Ciências das Computação, além da apresentação de alguns trabalhos científicos que demonstram a aplicação desse conceito e que estão, de alguma forma, alinhados com a proposta desse trabalho.

2.1 Ontologia

Segundo Castro (2008), a palavra "ontologia" foi criada por R. Goclenius para o seu Lexicon Philosophicum, publicado em 1613. Ela é resultado da junção de dois termos gregos, onta (entes) e logos (teoria, discurso, palavra). Ao pé da letra, ontologia significa, portanto, teoria dos entes. Ontologia seria, então, a teoria do ser enquanto tal. Sua função primeira seria a de estabelecer uma estratégia de organização dos seres e das coisas, sua relação com o mundo, a realidade e o conhecimento. Embora tenha suas bases na filosofia, é utilizada tanto na Ciência da Computação quanto na Ciência da Informação como possibilidade de melhora na definição de um domínio de conhecimento.

2.1.1 Ontologias e Ciências da Computação

No contexto das Ciências da Computação, Gruber (1993) define ontologia como uma especificação formal e explícita de uma conceitualização compartilhada. A ontologia utiliza o conceito de classes, atributos e relacionamentos. Essas informações reúnem dados sobre seus significados, restrições e aplicações lógicas existentes. Normalmente, são desenvolvidas numa linguagem que permite que a abstração da estrutura de dados e estratégias de implementação. Ainda segundo Gruber, devido à independência dos dados de nível inferior as ontologias são utilizadas para integrar bancos de dados heterogêneos, interoperabilidade entre sistemas diferentes e especificando interfaces para serviços independentes baseados no conhecimento. O autor ainda afirma que o processo de construção de uma ontologia é complexo, pois envolve a criação de

modelos semânticos ou descrições simplificadas da realidade de um dado domínio e, também, exige dos desenvolvedores conhecimentos técnicos em modelagem conceitual, em lógica formal e em alguns aspectos filosóficos.

2.1.2 Classificação de uma Ontologia

As Ontologias podem ser classificadas sob vários aspectos: de acordo com seu conteúdo e seu grau de formalidade.

Guarino (1998) classifica as ontologias de acordo com seu conteúdo:

- Ontologias Genéricas – descrevem conceitos gerais, tais como espaço, tempo, matéria, objeto, evento, ação, etc., que são independentes de um domínio particular.

- Ontologias de Domínio – descrevem um vocabulário relacionado a um domínio em particular.

- Ontologia de Tarefas – descrevem conceitos relacionados a tarefas ou atividades genéricas.

- Ontologias de Aplicação - descrevem conceitos que dependem tanto de um domínio específico como de uma tarefa específica, e geralmente é uma especialização de ambos.

Em se tratando do grau de formalidade, as ontologias são classificadas seguindo a notação de DING e ENGELS (2001) e GÓMEZ-PÉREZ (2004):

- Ontologias leves (light-weight), que são ontologias com pouco rigor formal, geralmente, composta de classes facilmente compreensíveis e de relações mais comuns entre estas classes, não incluindo relações especiais – por exemplo, relações lógicas - entre as classes e outros tipos de primitivas de representação – por exemplo, axiomas lógicos.

- Ontologias pesadas (heavy-weight), que se referem a ontologias com alto rigor formal, incluindo além das classes e relações comuns, relações especiais e alto grau de axiomatização. Em geral, essas ontologias são especificadas com um rigoroso grau matemático e formal, baseado nas linguagens lógicas, tal como a lógica de primeira ordem.

2.1.3 Metodologias para desenvolvimento de ontologias

Embora existam inúmeras pesquisas sobre o desenvolvimento de ontologias no campo da Computação sobre criação de ferramentas e formatos, não existe uma única forma de elaboração desse domínio. São vários os fatores que podem explicar a falta de metodologias mais concretas para essa elaboração:

> Mesmo com o crescimento notável das atividades de pesquisa sobre ontologias e de suas aplicações práticas nos últimos anos, algumas necessidades e problemas da área permanecem em aberto, talvez em função da imaturidade de ser uma área de pesquisa recente, a qual é denominada de engenharia ontológica. Dentre os problemas identificados, atualmente, na área de engenharia ontológica, aqueles que são objeto de estudo da presente pesquisa correspondem aos problemas que surgem ao longo do processo de construção de uma ontologia(...) Mendonça (2015).

Uma hipótese para essa não padronização de formato de criação de ontologia é justamente pela especificidade de cada caso. Uma ontologia criada para a área da saúde, por exemplo, não necessariamente terá a mesma eficácia em outro contexto. Elaborar uma metodologia para cada situação, também, se mostra inviável. As principais metodologias para desenvolvimento de ontologias segundo Mendonça (2015) são:

- Metodologia de Gruninger e Fox: Metodologia criada com base no desenvolvimento do projeto Toronto Virtual Enterprise (TOVE), cujo objetivo era o de criar um modelo de senso comum ou conhecimento compartilhado sobre empresas.
- Metodologia de Uschold e King: Tem como propósito principal descrever o conhecimento sobre domínios corporativos ou de negócios.
- Methontology: Que possibilita a construção de uma ontologia por reengenharia sobre outra ontologia, utilizando-se do conhecimento do domínio tratado.
- Método Kactus: Método recursivo derivado do projeto Kactus que permitiu a reutilização de conhecimento em sistemas de complexidade técnica, tal como o domínio de redes elétricas, e a construção de ontologias nesse domínio como suporte a tais sistemas.
- Método Sensus: Método derivado da ontologia Sensus, a qual foi desenvolvida pelo grupo Information Sciences Institute (ISI) com o propósito de ser usada para fins de processamento de linguagem natural.
- Método 101: Método concebido a partir da experiência no desenvolvimento de uma ontologia de vinhos e alimentos.
- Método CYC: Considera o conhecimento consensual do mundo e é indicada pelos autores na criação de ontologias para fundamentar sistemas inteligentes.

- On-to-Knowledge Methodology: desenvolvida para a construção de ontologias para aplicações de gestão do conhecimento, com o foco em Processo de Conhecimento (Knowledge Process) e em Conhecimento do Processo Meta (Knowledge Meta Process).
- UP for ONtology (UPON): Derivada e baseada no padrão de engenharia de software conhecido como Processo Unificado – do inglês Unified Software Development Process ou Unified Process (UP)– do qual foram derivadas de metodologias de software como o Rational Unified Process (RUP) e o Processo para Aplicativos extensíveis e Interativos (PRAXIS).
- Metodologia NeON: Metodologia para construção de redes ontológicas baseado em um desenvolvimento colaborativo e argumentativo de ontologias.
- Metodologia MFPFO: Tal metodologia é capaz de sugerir, automaticamente, anotações semanticamente relacionadas, baseadas no design e no repositório de construção.
- Ciclo de Vida de Schiessl e Bräscher: Embora não seja propriamente uma metodologia de construção de ontologias, o ciclo de vida ontológico, descrito por Schiessl e Bräscher (2012), inclui todas as etapas necessárias no processo de construção de ontologias, destacando o papel de cada etapa e as tarefas contidas em cada uma delas.

Na figura 1, as metodologias mencionadas por Mendonça (2015) são comparadas:

Figura 1 – Metodologias para desenvolvimento de ontologias

Metodologia	Descrição (objetivo e domínio)	Etapas (processos e atividades)
Metodologia de Gruninger e Fox – TOVE (GRUNINGER e FOX, 1995)	Criada no projeto *Toronto Virtual Enterprise (TOVE),* cujo objetivo era criar um modelo de senso comum compartilhado corporativo. **Domínio de aplicação:** Negócios (empresarial)	1. Elaboração de cenários de motivação 2. Questões de competência informal 3. Concepção da terminologia formal 4. Questões de competência formal 5. Especificação de axiomas formais; 6. Verificação de teoremas completos.
Metodologia de Uschold e King ENTERPRISE (USCHOLD e KING, 1995)	Desenvolvido com base na prática da construção da ontologia *Enterprise Ontology*, que descreve conhecimento nodomínio corporativos. **Domínio de aplicação:** Negócios (empresarial)	1. Propósito, grau de formalismo e usuário 2. Construção da ontologia em três etapas: a) conceitualização b) implementação c) integração com ontologias já existentes; 3. Avaliação 4. Documentação
Methontology (GÓMEZ- PEREZ, FERNANDEZ-LOPES e VICENTE, 1996)	Possibilita a construção de ontologias por reengenharia, utilizando-se do conhecimentode domínio. **Domínio de aplicação:** Diversos.	1. Especificação: 2. Aquisição de Conhecimento 3. Conceitualização 4. Integração 5. Implementação 6. Avaliação 7. Documentação
Método *Kactus* (BERNARAS, LARESGOTTI, CORERA, 1996)	Método derivado do projeto Kactus que permite reutilização de conhecimento em sistemas técnicos, tal comoredes elétricas. **Domínio de aplicação:** Sistemas de complexidade técnica.	1. Lista de necessidades ou requisitos 2. Identificação de termos relevantes 3. Criação de modelo preliminar 4. Estruturação e refinamento 5. Integração e reutilização.
Método *Sensus* (SWARTOUT et al., 1996).	Método derivado da ontologia*Sensus*, desenvolvida para de processamento de linguagem natural. **Domínio de aplicação:** Diversos.	1. Identificar termos-chave do domínio; 2. Ligar os termos-chave à ontologia *Sensus*; 3. Adicionar novos termos para o domínio; 4. Adicionar subárvores completas.
Método 101 (NOY e GUINESS, 2001)	Método usado no desenvolvimento de exemplospráticos utilizando o editor deontologias Protégé. **Domínio de aplicação:** Diversos.	1. Determinar escopo 2. Considerar o reuso de termos 3. Enumerar termos 4. Definir classes 5. Organizar as classes em uma taxonomia 6. Definir propriedades e suas restrições 7. Adicionar valores de instâncias
Método CYC (REED, LENAT, 2002)	Método usado na construção da ontologia CYC, que buscaabranger o considera o conhecimento consensual do mundo. **Domínio de aplicação:** Diversos	1. Extração manual do conhecimento 2. Extração auxiliada por computador 3. Desenvolvimento de representação 4. Representação do conhecimento de diferentesdomínios usando primitivas.

On-to-Knowledge Methodology (OTKM) (SURE, STAAB e STUBER, 2003)	Metodologia desenvolvida para a construção de ontologias para aplicações degestão do conhecimento **Domínio de aplicação:** Gestão do conhecimento empresarial.	1. Estudo de viabilidade 2. *Kickoff* 3. Refinamento 4. Avaliação 5. Aplicação e evolução 6. Geração de conhecimento 7. Captura de conhecimento 8. Recuperação e acesso
Metodologia *UP for ONtology* (UPON) (DE NICOLA, MISSIKOFF e NAVIGLI, 2009)	Metodologia de construção de ontologias derivada em padrãode engenharia de software Processo Unificado **Domínio de aplicação:** Negócios (*e-bussiness*).	1. Workflow de Requisitos 2. Workflow de Análise 3. Workflow de Desenvolvimento 4. Workflow de Implementação 5. Workflow de Teste
Metodologia *NeON* (SUARÉZ-FIGUEROA, 2010)	Metodologia para construçãode redes ontológicas baseadoem um desenvolvimento colaborativo **Domínio de aplicação:** Diversos.	1. Gerenciamento de processos e atividades 2. Desenvolvimento orientado de processos e atividades 3. Suporte aos processos e atividades
Metodologia *MFPFO* (LIM, LIU eLEE, 2011)	Metodologia de construção de ontologia multi-facetada, anotada semanticamente, para a modelagem de uma famíliade produtos. **Domínio de Aplicação:** Domínios que possuem uma família de produtos	1. Construção de uma taxonomia da família de produtos; 2. Extração de entidades; 3. Identificação do conceito e geração da unidade facetada; 4. Anotação semântica e modelagem faceta; 5. Construção de uma ontologia de família de produtos multi-facetada e anotada semanticamente; 6. Avaliação e validação da ontologia.
Ciclo de Vida de Schiessl e Bräscher (2011)	Embora não seja uma metodologia, o ciclo de vida ontológico inclui todas as etapas necessárias no processode construção de ontologias **Domínio de Aplicação:** Diversos.	1. Especificação 2. Conceitualização 3. Formalização 4. Aplicação 5. Manutenção 6. Atividades paralelas: aquisição de conhecimento, avaliação, documentação.

Fonte: Mendonça (2015)

Para este trabalho, considerando exemplos de aplicação já consolidados em trabalhos anteriores, a proposta para a criação de uma Ontologia de domínio do Grupo de Estudos Olímpicos (OntOlympic) se baseia no modelo desenvolvido pelo Professor Fabrício Martins Mendonça, da Universidade Federal de Juiz de Fora, em Minas Gerais, a Ontoforinfoscience.

2.1.4 Metodologia Ontoforinfoscience

De acordo com Mendonça (2015), a Ontoforinfoscience tem como principal diferencial em relação às outras metodologias disponíveis, um detalhamento das atividades necessárias do ciclo de desenvolvimento ontológico, a fim de auxiliar especialistas em organização do conhecimento, incluindo cientistas da informação, a superar problemas relativos ao jargão técnico e às questões lógicas e filosóficas que envolvem a construção de ontologias. A metodologia Ontoforinfoscience representa, portanto, uma iniciativa em direção a maior entendimento de termos e detalhes técnicos (lógicos e filosóficos) do processo de desenvolvimento de ontologias por parte dos cientistas da informação. Ainda que apresente uma característica particular, a metodologia Ontoforinfoscience também se baseou em algumas etapas de metodologias já existentes, especificamente a Methontology, o método 101 Method e metodologia NeOn, a fim de reutilizar algumas etapas e suprir as limitações presentes em cada uma delas: detalham-se as etapas e passos reutilizados, além de realizar adaptações necessárias para obter uma linguagem apropriada a todos desenvolvedores de ontologia.

2.1.4.1 Ontoforinfoscience: Etapas

A metodologia Ontoforinfoscience é composta por nove etapas. Uma primeira etapa de avaliação da necessidade da construção de uma ontologia. A partir dela (caso seja comprovada a necessidade) o processo segue para as demais. Caso essa primeira etapa não confirme a necessidade de construção de uma ontologia, recomenda-se a utilização de outros instrumentos terminológicos, como tesauros, vocabulários controlados ou ainda, taxonomias.

2.1.4.1.1 Pré-etapa (Etapa 0)

Nesse momento, é avaliado se a construção de uma ontologia é necessária. Recomenda-se criar a ontologia quando for preciso uma indexação e recuperação de informação em um contexto dinâmico para descrição de recursos de um dado domínio do conhecimento, além da possibilidade de uma inferência automatizada. Além disso, considera-se como requisito para criação de uma ontologia, representar aspectos e objetos do mundo real, a necessidade de mais relações entre os objetos de um domínio além daquelas contidas em vocabulários controlados e o uso de formalismos lógicos para representação da informação e inferência automatizada segundo Mendonça (2015).

2.1.4.1.2 Etapa 01 - Especificação

Na etapa 1 realiza-se a especificação da ontologia através do modelo de especificação, o qual deve conter informações sobre o domínio e escopo, propósito geral, classes de usuários que representam o público-alvo da ontologia, cenários de aplicação para uso da ontologia e o grau de formalidade. Estabelece-se também o tipo da ontologia: ontologia de alto ou médio

nível, de domínio ou de tarefa; ontologia leve ou pesada. Por fim, delimita-se o escopo de cobertura da ontologia descrevendo o ponto de partida, o limite do domínio coberto e questões de competência.

2.1.4.1.3 Etapa 02 - Aquisição e Extração de conhecimento

Logo após a etapa de especificação da ontologia, temos a etapa de aquisição e extração de conhecimento do domínio. São consideradas nesse momento, todas as referências possíveis para a representação desse domínio.

Esse material selecionado, formará um conjunto de referências da ontologia. Mendonça (2015), aponta quais são os métodos e técnicas que podem ser utilizados para esse passo:

- análise informal de textos em materiais de referência do domínio: permite o estudo e entendimento dos conceitos nos materiais utilizados;
- análise formal de textos nos documentos de referência do domínio: consiste na identificação de estruturas textuais do domínio, tais como definição e afirmação, e o tipo de conhecimento que tais estruturas podem representar na ontologia: conceitos (classes), propriedades, instâncias, relações, etc.
- entrevistas estruturadas, semi-estruturadas ou não- estruturadas com especialistas da área: possibilitam o entendimento de conceitos do domínio por parte do desenvolvedor e a construção de árvores de classificação de conceitos do domínio.
- *brainstorming* ou grupos focais com especialistas da área: é uma técnica mais colaborativa que a realização de entrevistas para a compreensão de conceitos do domínio, porém pode se tornar ineficiente quando não se obtém consenso sobre os termos da área, necessitando de intervenção contínua do mediador desta atividade.
- emprego de técnicas estatísticas e/ou métodos linguísticos que possibilitem a extração de conceitos e relações relevantes do domínio a partir da análise dos textos dos materiais de referência utilizados.
- uso de ferramentas automatizadas que realizam a extração de conceitos e relações relevantes do domínio. De maneira geral, tais ferramentas são usadas a partir de uma abordagem semi-automática, tal que o desenvolvedor ou especialista do domínio avaliem e validem os termos candidatos extraídos de maneira automática.

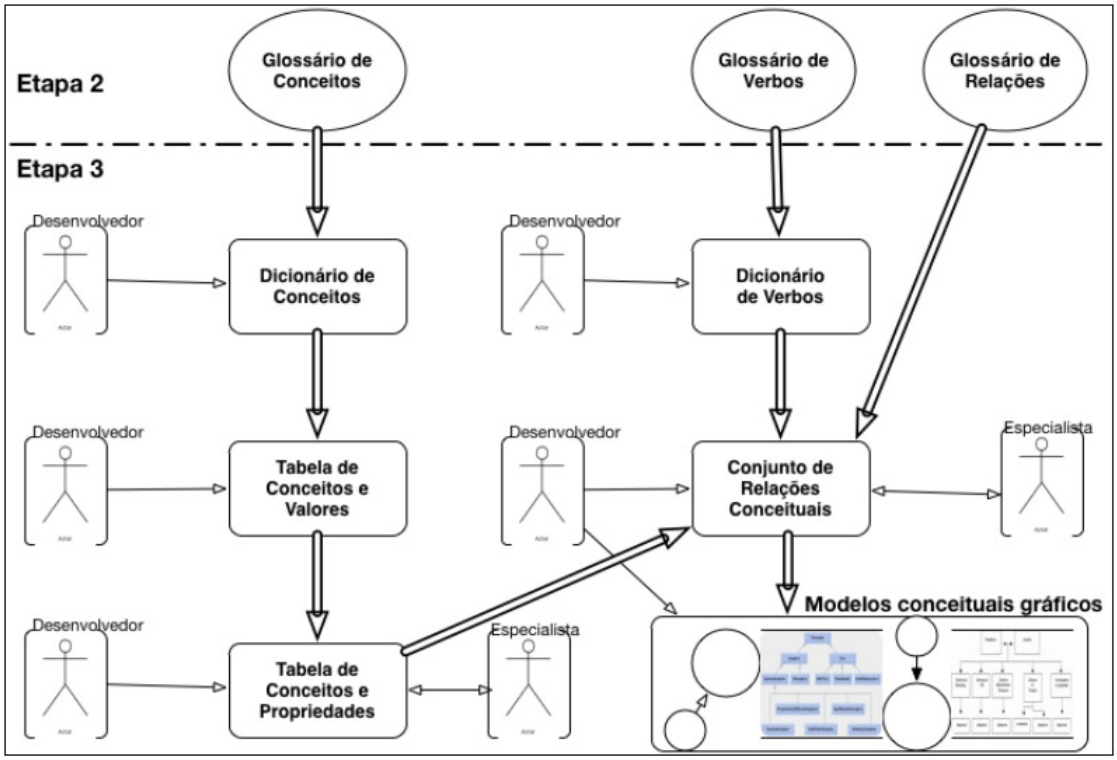
Após a realização dessa extração, o seu resultado é a formulação de glossários, divididos em:

- Glossário de Conceitos: nesse glossário, são considerados os substantivos, pronomes e adjetivos. São considerados conceitos do domínio e prováveis classes na ontologia;
- Glossário de Verbos: aqui, são extraídos dos textos de referência todos os verbos relevantes, que são candidatos a se tornarem relações na ontologia;
- Glossário de Relações: conjunto formado por todos os padrões de relacionamento extraídos dos textos na forma <nome> <verbo> <nome>, tal que os nomes podem ser substantivos, pronomes ou adjetivos e o verbo representa um relacionamento existente entre os nomes.

2.1.4.1.4 Etapa 3 - Conceitualização

Nesse momento, todo o levantamento realizado na etapa de aquisição do conhecimento é utilizado para criar a estrutura conceitual da ontologia. A partir dos glossários elaborados na etapa de aquisição, são criados dicionários dos conceitos e dos verbos (figura 2), além de uma tabela de relações. A partir dos dicionários de conceitos serão conhecidas as classes principais dessa ontologia. A partir dos verbos, provavelmente surgirão as relações entre as classes.

Figura 2 – Etapa 3: Conceitualização



Fonte: Mendonça (2015)

2.1.4.1.5 Etapa 4 - Fundamentação

A metodologia Ontoforinfoscience traz como sugestão nessa etapa a especificação de uma fundamentação teórica filosófica que orientaria o desenvolvimento. Essa especificação deve incluir a abordagem filosófica adotada como base para a construção ontológica e a possibilidade de uso de uma ontologia de fundamentação nesse processo de construção. Portanto, o desenvolvedor deverá escolher uma abordagem filosofica ontológica, além de selecionar outras ontologias já estabelecidas como base para a ontologia a ser desenvolvida. Entretanto, em algumas situações, esse passo pode ser considerado opcional. No caso de ontologias menores, leves e de domínio, pode não ser necessário essa fundamentação prévia.

> Em um projeto ontológico de menor porte, onde se deseja construir uma ontologia leve de domínio ou mesmo uma sub-ontologia pode não ser necessária a realização desta etapa de fundamentação. Entretanto, na grande maioria dos projetos ontológicos, essa é uma etapa essencialmente importante para cumprir com os propósitos da ontologia desenvolvida. (Mendonça (2015))

2.1.4.1.6 Etapa 5 - Formalização

Para Mendonça (2015) essa é a etapa que produz uma descrição formal do domínio sob estudo baseada na conceitualização deste domínio realizada anteriormente na etapa 3 (três). O conhecimento do domínio, tratado anteriormente apenas a nível conceitual, passa a ser tratado a nível ontológico-formal, o que implica em uma série de adaptações nas estruturas conceituais de forma a atender as restrições ontológicas e formais desta etapa.

2.1.4.1.7 Etapa 6 - Avaliação

Segundo a metodologia Ontoforinfoscience, vários critérios podem ser considerados para a validação da ontologia. A metodologia oferece, inclusive, um modelo que pode ser utilizado como orientador para essa fase. Dois critérios básicos são avaliados nesse momento:a validação e a verificação da ontologia.

O critério de validação, verifica se a ontologia realmente atende a realidade que ela representa. Ou seja, se ela conceitualmente está correta. Esse passo pode ser realizado, por exemplo, com a ajuda de especialistas de domínio.

Entende-se verificação como o funcionamento da mesma, a partir do momento em que ela é criada e interpretada computacionalmente. Se ela atende as restrições, seus princípios básicos ontológicos e se não existe nenhum erro em sua concepção. Essa etapa pode ser realizada utilizando, por exemplo, o reasoner (classificador) das ferramentas de elaboração de ontologia. Ao ativar esses classificadores, os possíveis erros de restrições, por exemplo, aparecem, indicando necessidade de revisão.

2.1.4.1.8 Etapa 7 - Documentação

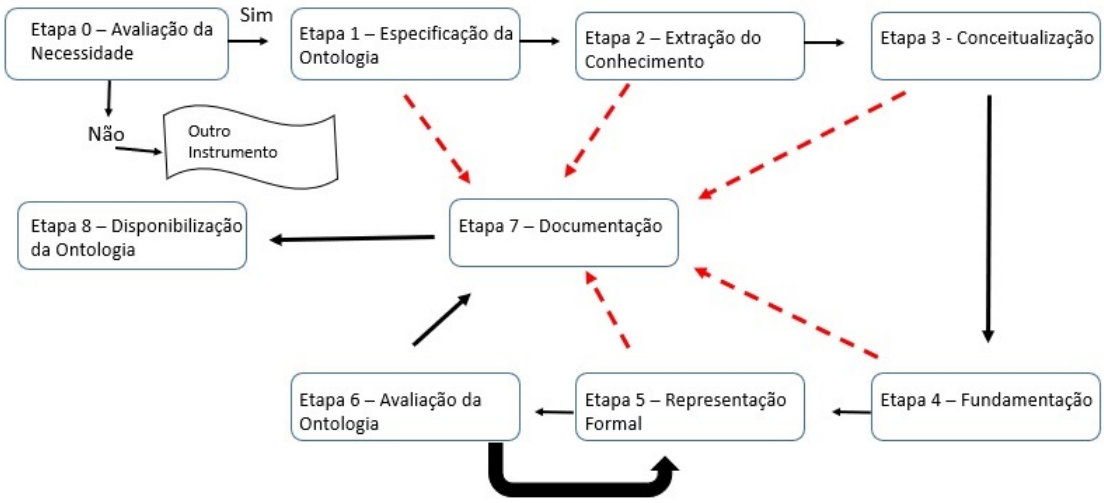
Embora seja apontado como uma etapa em separado, essa fase da elaboração da ontologia que fala da documentação normalmente já é feita ao longo do processo. Nessa etapa, são apontados alguns modelos de como essa documentação pode ser formalizada, mas preferencialmente, em cada etapa anterior a sua documentação deverá ser elaborada, visando formalizar a ontologia desenvolvida.

2.1.4.1.9 Etapa 8 - Disponibilização da Ontologia

Na última etapa do processo, a ontologia deverá ser disponibilizada em meio eletrônico, para fácil acesso de outros usuários. Com isso, essa ontologia poderá ser utilizada como base para o desenvolvimento de novas ontologias.

Temos, portanto, as oito (ou nove, considerando a etapa 0) etapas para a criação da ontologia através da metodologia Ontoforinfoscience (Figura 3).

Figura 3 – Esquema representando a metodologia Ontoforinfoscience



Fonte: Elaborado pelo autor

2.2 Anotação Semântica

Segundo Bontcheva K. (2011), o processo de vinculação de modelos semânticos em conjunto com a linguagem natural é referido como anotação semântica. No caso da anotação via ontologia, esse processo pode ser caracterizado como a criação dinâmica de inter-relações entre ontologias e documentos não estruturados e semiestruturados de forma bidirecional. Do ponto de vista tecnológico, a anotação semântica insere nos textos todas as menções de conceitos da ontologia (ou seja, classes, instâncias, propriedades e relações), por meio de metadados referentes aos seus Uniform Resource Identifier (URIs) na ontologia. Abordagens que apenas aprimoram a ontologia com novas instâncias derivadas dos textos são normalmente referidas como população de ontologia.

2.3 Mineração de Textos

Quando os dados disponíveis são textos escritos em língua natural, ou seja, dados não estruturados, o processo de extração de conhecimento é chamado de mineração de textos (do inglês *text mining*). Para Rezende, Marcacini e Moura (2011), métodos não supervisionados para extração e organização de conhecimento recebem grande atenção na literatura, uma vez que não exigem conhecimento prévio a respeito das coleções textuais a serem exploradas. O processo pode ser dividido em três fases principais: Pré-Processamento dos Documentos, Extração de Padrões e Avaliação do Conhecimento. Rezende, Marcacini e Moura (2011) apontam ainda a importância do pré-processamento na mineração de textos. Nessa etapa, os dados textuais são padronizados e representados de forma estruturada e concisa, em um formato adequado para extração do conhecimento. É usual a normalização do textos, que pode incluir remoção de palavras irrelevantes, stemming, lematização, tokenização e vetorização, entre outras tarefas. Uma abordagem bastante explorada em pré-processamento é utilizar o índice TF-IDF (Term Frequency-Inverse Document Frequency) para representar os documentos da coleção analisada. O TF-IDF é uma medida estatística destinada a medir o grau de importância de uma palavra para um conjunto de documentos (Salton e Yang, 1973, APUD Cavalcanti (2018)). Um termo, ao aparecer repetidamente em muitos documentos, é considerado menos importante, tendendo a zero. TF-IDF combina a frequência dos termos (TF) e a relevância do termo para todo o conjunto de textos (IDF). Cavalcanti (2018) apresenta a seguinte fórmula para representar esse modelo, onde a equação 2.1 representa a frequência de termos, a equação 2.2 representa a relevância de um termo para uma coleção, e a equação 2.3 a multiplicação das duas anteriores.

$$TF = (\frac{\text{número de vezes que um termo aparece em dada sentença}}{\text{número total de termos presentes na sentença}}) \tag{2.1}$$

$$IDF = 1 + log_e(\frac{\text{número total de sentenças}}{\text{quantidade de sentenças que apresentam determinados termo}}) \tag{2.2}$$

$$TF - IDF = TF \times IDF \tag{2.3}$$

Na etapa seguinte, de Extração de Padrões, métodos de agrupamento de textos, por exemplo, podem ser utilizados para a organização de coleções textuais de maneira não supervisionada. Em agrupamentos, o objetivo é organizar um conjunto de documentos em grupos, em que documentos de um mesmo grupo são similares entre si, mas dissimilares em relação aos documentos de outros grupos (Rezende, Marcacini e Moura (2011)). Dentre os diversos métodos de agrupamento de dados, destacam-se em mineração de textos como os particionais (K-means) e hierárquicos (Aglomerativos). No agrupamento particional, o conjunto de objetos é dividido interativamente em k grupos, sendo que o k é normalmente indicado pelo usuário previamente (CORRÊA, MARCACINI e REZENDE (2012)). Linden (2009) afirma que o agrupamento hierarquico criam uma hierarquia de relacionamentos entre os elementos. O agrupamento hierarquico é dividido em duas partes, segundo CORRÊA, MARCACINI e REZENDE (2012): agrupamento hierárquico aglomerativo (onde cada elemento pertence a um grupo, e depois se juntam a outros grupos até formar um único grupo novamente), e os divisivos (onde se tem um grupo único inicialmente e grupos menores se formam até que os elementos se tornem únicos).

Para agrupamento de texto, algumas medidas de qualidade segundo Godois (2018): índice de Dunn, Coeficiente de Silhouetta, Davies-Bouldin entre outros. Essas medidas são classificadas como internas, ou seja, aqueles que não precisam de informações além dos próprios agrupamentos. Essas medidas visam demonstrar através de uma fórmula matemática, em quais situações os agrupamentos podem ser melhor representados.

A figura 4 apresenta um resumo desses índices:

Figura 4 – Medidas de qualidade para agrupamentos

Nome do índice	*Descrição*
Silhouette	Esse índice determina a qualidade dos agrupamentos baseado na proximidade entre os objetos de um determinado grupo, além da proximidade desses objetos ao grupo que está mais próximo. Resulta em valores que variam entre [-1, 1]. Quanto mais próximo de 1 melhor o objeto está alocado no grupo.
Davies-Bouldin	O cálculo do índice está em função da razão entre a soma da dispersão interna dos agrupamentos e a distância entre dos referidos agrupamentos. Logo, como resultados desejáveis se tem que menores valores desse índice correspondem a agrupamentos compactos e com os centroides distantes entre si.
Dunn	O índice é calculado pela razão entre a menor distância intergrupo e a maior distância intragrupo. O resultado varia no intervalo (0, 1), de tal forma que quanto maior o valor resultante mais compactos e bem separados são os grupos.

Fonte: Godois (2018) - Adaptado pelo autor

É importante ressaltar que, apenas o índice não é suficiente para um resultado final do processo. A participação dos especialistas da área, analisando semanticamente o que cada grupo pode representar, é fundamental no processo de descoberta de conhecimento.

2.4 Considerações Finais

Nesse capítulo, foram apresentados os principais conceitos relacionados à ontologia, anotação semântica e mineração de textos. São apresentadas informações relacionadas às metodologias para criação de ontologia, aplicação da anotação semântica e da mineração de textos. O próximo capítulo abordará alguns dos trabalhos relacionados a esses temas.

CAPÍTULO

3

TRABALHOS RELACIONADOS

Damasceno, Ribeiro e Reategui (2011) demonstram em seu estudo como a integração da ferramenta de mineração de textos com uma ontologia de domínio, junto da catalogação CID, pode relacionar semanticamente documentos, possibilitando uma busca, navegação e acesso mais ágil.

Pereira (2014) traz em sua dissertação de Mestrado uma abordagem de anotação semântica automática em documentos históricos do século XIX. Foi construída uma ontologia de domínio (Ontologia Instrumental Linguistico) que apoiou o processo de anotação semântica automática. Os resultados apresentaram alto grau de coincidência, comprovando a eficácia da abordagem de anotação semântica automatica.

Santos (2016) desenvolveu em sua dissertação de Mestrado a ontologia OntoFootballForNewspapers. Essa ontologia, desenvolvida com a metodologia Ontoforinfoscience, visa apoiar a recuperação da informação relacionada a futebol em jornais digitais. Uma dificuldade nesse trabalho em especial é o fato de que as notícias sobre esse assunto teve, ao longo do tempo, muitas mudanças linguísticas. Dificulta o acesso via busca simples, pois acaba limitando os resultados aos termos mais usuais de uma determinada época.Como resultado, o trabalho concluiu que a inclusão de ontologias ao esquema de buscas para essas notícias do futebol, trouxe resultados menos exaustivos do que as buscas realizadas sem a aplicação da ontologia.

Arruda (2017), propõe um método semântico baseado em ontologia (SOM4SImD) para detectar similaridade entre documentos no contexto da educação especial. Os resultados mostraram que o SOM4SImD é eficaz na obtenção de similaridade entre documentos. A título de comparação, esse método teve um índice de precisão de 0,96, contra 0,71 de outro trabalho com características equivalentes.

Godois (2018) traz em sua dissertação a avaliação de um algoritmo de clustering (agrupamento) baseado em características de agentes, que detecta o número de grupos para um determinado conjunto de dados.

RUBIO *et al.* (2019) em artigo publicado em revista cientifica, faz uma análise das entrevistas dos atletas olímpicos brasileiros com mineração de textos. Há a hipótese de que os resultados corroboram para o estudo de associações entre o conteúdo mais frequentemente presente em narrativas e o cruzamento de dados com luz às teorias psicológicas de traços de personalidade. Ao final, verificou-se a relação entre adjetivos utilizados na narrativa com traços de personalidade.O trabalho servirá de base para o desenvolvimento de métodos baseados em mineração de textos para capturar traços de personalidade, utilizando o modelo dos Cinco Grandes Fatores de Personalidade.

Coelho (2022) apresenta uma nova ontologia para anotação semântica, que possibilita a anotação de conteúdo (mídia normal) a ser usado por mídias educacionais enativas. A ontologia proposta reutiliza e estende os padrões internacionais para anotação de mídia e promove a interoperabilidade com os modelos existentes.

3.1 Considerações Finais

Nesse capítulo, foram apresentados alguns dos trabalhos relacionados à ontologia, anotação semântica e mineração de textos. O trabalho de Arruda (2017), traz a abordagem mais próxima a esse trabalho, pois trata da busca de similaridade entre documentos na área de educação especial utilizando ontologia e anotação semântica. Já Santos (2016), traz conceitos relacionados à criação de uma nova ontologia de domínio, que também foi realizada nesse trabalho, enquanto Coelho (2022) apresenta uma nova ontologia a ser aplicada com anotação semântica. Pereira (2014) contribui com o processo de anotação semântica, enquanto Godois (2018) e Damasceno, Ribeiro e Reategui (2011) trazem contribuições importantes no campo da mineração de textos. O artigo de RUBIO *et al.* (2019), apresenta um estudo diretamente ligado a entrevista dos atletas olímpicos utilizando mineração de texto. Todos os trabalhos, portanto, de alguma maneira trouxeram contribuições o processo de anotação semântica em entrevistas utilizando ontologia de domínio, com objetivo de aplicar essas entrevistas num processo de mineração de textos. O próximo capítulo demonstra como a ontologia OntOlympic foi desenvolvida, utilizando a metodologia Ontoforinfoscience.

CAPÍTULO

4

ONTOLOGIA ONTOLYMPIC

Neste trabalho, é proposta uma ontologia de domínio, chamada OntOlympic, criada para apoiar trabalhos do Grupo de Estudos Olímpicos da Universidade de São Paulo. Essa ontologia pretende atender os principais temas trabalhados nesse grupo de estudos. Foi realizada uma pesquisa no grupo, visando levantar os temas que já foram objetos de pesquisa, os que estão no processo e possíveis temas que podem vir a ser investigados no futuro. O grupo possui diversos trabalhos publicados sobre esses temas, o que facilitou o processo de aquisição de conhecimento. A OntOlympic foi desenvolvida utilizando a metodologia Ontoforinfoscience, como descrito a seguir.

4.1 Etapa 0: Avaliação da necessidade da Ontologia

Mendonça (2015) afirma que, a condição principal para a criação de uma ontologia é a possibilidade de descrever os recursos de um domínio objetivando a recuperação de informações em um contexto dinâmico. A ontologia deve representar essa realidade e ser interpretada por humanos e máquinas. Sendo assim, a criação da ontologia OntOlympic se justifica, uma vez que será utilizada num contexto de recuperação da informação e de um domínio específico com vistas a aproveitar os recursos que as máquinas podem oferecer, como a inferência automática. Consideramos inferência a capacidade que uma ferramenta tem de concluir determinada situação a partir das regras que foram definidas. Por exemplo: considerando que São Carlos etá associado ao estados de São Paulo. E o estado de São Paulo está associado ao país Brasil. Através dessa ligação, a inferência consegue definir que São Carlos está associado ao pais Brasil. A ontologia OntOlympic permitirá que deduções assim sejam realizadas de maneira automática.

4.2 Etapa 01: Especificação da Ontologia

Seguindo o modelo da metodologia, foram definidos para a OntOlympic (Figura 5):

Figura 5 – Escopo da ontologia OntOlympic

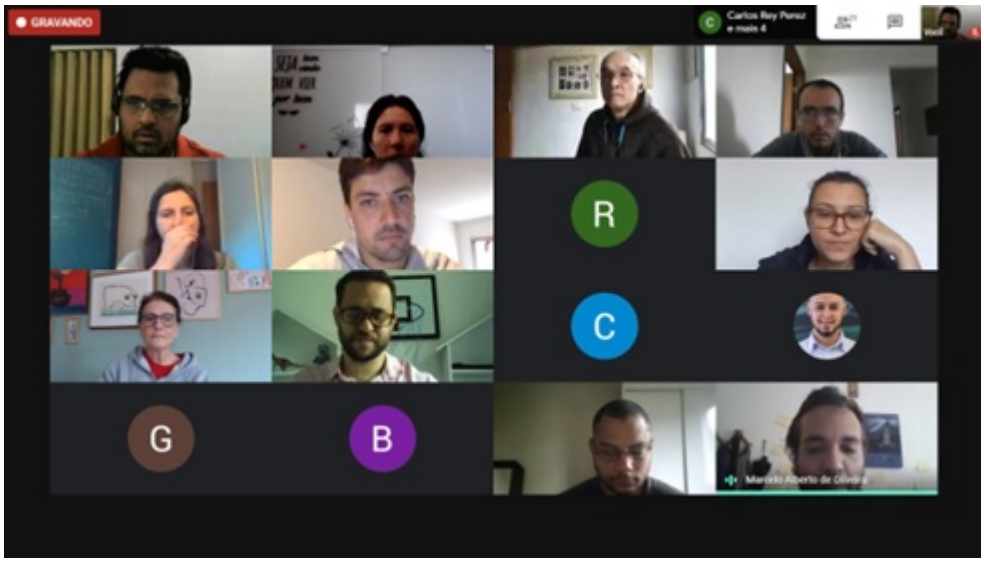
Domínio/Escopo Geral
A OntOlympic é uma ontologia da área esportiva. Seu domínio representa o conhecimento relativo aos atletas olímpicos brasileiros, adquirido por meio de entrevistas ou documentos biográficos históricos. Os documentos, representam uma amostra absolutamente especial, visto que os atletas olímpicos representam uma parcela mínima da população brasileira, porém, altamente especializada e única em seu contexto. Existe a hipótese de que esses documentos podem apresentar informações preciosas para pesquisas futuras dentro do meio acadêmico, visto sua variedade: pesquisadores das áreas de história do esporte, psicologia do esporte, educação física, medicina, jornalismo, entre outras, configurando um verdadeiro trabalho interdisciplinar.
Propósito Geral
A OntOlympic tem como objetivo geral ser um instrumento de auxílio aos pesquisadores da Universidade de São Paulo, mais precisamente do Grupo de Estudos Olímpicos da Faculdade de Educação. Existem inúmeras hipóteses que poderão ser comprovadas através dos documentos armazenados. Além desses pesquisadores, o público em geral também pode se beneficiar dessa ontologia, visto que alguns dados poderão ser disponibilizados em uma plataforma na internet. Conhecer o domínio desse universo é fundamental para auxiliar no processo de busca dentro do acervo.
Classes de usuários
A ontologia OntOlympic é destinada, especialmente, para os pesquisadores do Grupo de Estudos Olímpicos da Universidade de São Paulo. O grupo engloba profissionais da Educação Física, Psicologia, História e Sociologia do esporte, Geografia, Tecnologia da Informação, entre outros. Esses profissionais atuam na pesquisa, nos mais variados temas relacionados a sua área de atuação. O acesso aos documentos, de maneira objetiva e simplificada, permitirá que os esses pesquisadores tenham maior facilidade no uso e análise dos dados, trazendo respostas mais precisas às suas respectivas pesquisas.
Tipo da ontologia
A OntOlympic é classificada como ontologia de domínio (Grupo de Estudos Olímpicos da Faculdade de Educação da USP.
Grau de formalidade
A OntOlympic é uma ontologia leve, pois possui menos rigor formal e cujo objetivo é possibilitar a integração entre sistemas computacionais
Delimitação do escopo de cobertura
O ponto de partida da OntOlympic, são entidades do mundo real, correspondente ao domínio determinado (Grupo de Estudos Olímpicos da Faculdade de Educação da USP). As classes e subclasses correspondem aos temas tratados até o momento pelo grupo, ou temas já diagnosticados e que podem ser tratados no futuro. As classes principais são: Pessoa (representando os atores do processo), Jogos Olímpicos, Lugar, Saúde, Raça, Gênero, Transição, Doping, E-Sports, Militar.
Questões de competência:
Q1. Quais as lesões mais comuns são citadas pelos atletas? Q2. Quais as formações acadêmicas mais citadas pelos atletas? Q3. Quais os lugares mais citados pelos atletas olímpicos? Q4. Os professores são citados pelos atletas durante a entrevista? Q5. Que tipos de dor são mais citadas pelos atletas nas entrevistas? Q6. Que problemas de ordem de saúde mental são citados pelos atletas? Q7. Os atletas citam questões sobre preconceito (racismo, machismo, lgbtqfobia)? Q8. Que membros da família são os mais citados pelos atletas nas entrevistas? Q9. Os atletas citam seus técnicos nas entrevistas? Q10. Que problemas relacionados ao peso os atletas citam nas entrevistas?

Fonte: Elaborado pelo autor

4.3 Etapa 02: Aquisição e extração de conhecimento

Para o desenvolvimento da OntOlympic, foram utilizadas as seguintes estratégias para aquisição e extração do conhecimento: análise formal de textos de referência do domínio, entrevistas não-estruturadas com os especialistas da área, uso de ferramentas automatizadas para auxiliar no processo de decisão dos termos/temas/assuntos relevantes do GEO. Além dessas, que estão no escopo do desenvolvimento da ontologia, para este trabalho houve a participação do autor nas reuniões semanais do grupo (Figura 6).

Figura 6 – Reunião online semanal do Grupo de Estudos Olímpicos da USP



Fonte: Elaborado pelo autor

4.3.1 Análise Formal dos textos

Foram consideradas para a análise os seguintes textos, todos eles produzidos em algum momento por integrantes do grupo de estudos olímpicos.

Teses de Doutorado

- Luciane Maria Micheletti Tonon. Olímpicos e Paralímpicos: separados por um instante Retratos biográficos dos instantes significativos de atletas que transitaram entre os Movimentos Olímpico e Paralímpico. 2022

- Neilton Ferreira Junior. Olimpismo negro: uma antologia das resistências ao racismo no esporte, por atletas olímpicos brasileiros. 2021.

- William Douglas de Almeida. Brasileiros, por que não? Trajetória e identidade dos migrantes internacionais no esporte olímpico do Brasil. 2020.

- Rafael Campos Veloso. Trajetos entre alvoradas e crepúsculos: atleta e as muitas faces do mito do herói. 2021.
- Natália Kohatsu Quintilio. Das vivências às experiências significativas: os valores olímpicos como mobilizadores das habilidades socioemocionais por meio do esporte educacional. 2019.
- Dhenis Rosina. Entre narrativas, fragmentos e estilhas: construções de atletas brasileiros sobre os Jogos Olímpicos do México de 1968. 2018.
- Carlos Rey Perez. O entendimento de valores olímpicos por atletas olímpicos brasileiros. 2017.
- Ivan Sant'Ana Rabelo. Investigação de traços de personalidade em atletas brasileiros: análise da adequação de uma ferramenta de avaliação psicológica. 2013
- Sérgio Settani Giglio. Representações do futebol nos Jogos Olímpicos e na Copa do Mundo. 2012
- Alexandre Velly Nunes. A influência da imigração japonesa no desenvolvimento do judô brasileiro: uma genealogia dos atletas brasileiros medalhistas em Jogos Olímpicos e Campeonatos Mundiais. 2011.

Dissertações de Mestrado

- Maria Alice Zimmermann. O professor inesquecível nas narrativas de atletas olímpicos brasileiros. 2019.
- Julia Frias Amato. Aventura do Herói nas Histórias de Vida de Mulheres Olímpicas Brasileiras: a partida como primeiro estágio da Jornada Mitológica. 2018.
- Gabriela de Carvalho Monteiro Gonçalves. O significado da dor em atletas de ginástica rítmica. 2017.
- Neilton de Sousa Ferreira Junior. A transição de carreira dos bicampeões mundiais de basquetebol: uma análise com base em narrativas biográficas. 2014.
- David Alves de Souza Lima. Técnico-Mestre e Atleta-Herói. Leitura Simbólica dos Mitos de Quíron e do Herói entre técnicos de voleibol, 2012
- Paulo Henrique do Nascimento. Mulheres no pódio: as histórias de vida das primeiras medalhistas olímpicas brasileiras. 2012
- Danilo Luis Rodrigues Lemos. A história social do movimento olímpico brasileiro no início do século XX. 2008.

- Raoni Perrucci Toledo Machado. Esporte e religião no imaginário da Grécia Antiga. 2006.

Livros publicados

- RUBIO, K.. Atletas Olímpicos Brasileiros. 1. ed. São Paulo: SESI-SP Editora, 2015. v. 1. 648p.
- RUBIO, K.. Psicologia, Esporte e Valores olímpicos. 1. ed. São Paulo: Casa do Psicólogo, 2012. v. 1. 250p
- RUBIO, K.; QUINTILIO, N. K.; MARCONI, J. R.. Missão Valores Olímpicos. 1. ed. São Paulo: Laços, 2016. v. 1. 30p.
- RUBIO, K.. Narrativas biográficas: da busca à construção de um método. 1. ed. São Paulo: Editora Laços, 2016. v. 1. 287p.
- RUBIO, K.. Do pós ao neo olimpismo: esporte e movimento olímpico no século XXI. 1. ed. São Paulo: Laços, 2019. v. 1. 382p.
- RUBIO, K.. Esporte e Mito. 1. ed. São Paulo: Laços, 2018. v. 1. 206p .
- RUBIO, K.. Atletas do Brasil Olímpico. 1. ed. São Paulo: Kazuá, 2013. v. 1. 297p.
- RUBIO, K.. As mulheres e o esporte olímpico brasileiro. 1. ed. São Paulo: Casa do Psicólogo, 2011. v. 1. 284p .

Artigos publicados

- RUBIO, K.. Jogos Olímpicos da Era Moderna: uma proposta de periodização. Revista Brasileira de Educação Física e Esporte (Impresso), v. 24, p. 55-68, 2010.
- FRANCISCO, W. V. ; SANTOS, U. F. ; RUBIO, K. . O respeito ao não-dito nas narrativas de atletas lgbtqia+ do esporte olímpico. Olimpianos - Journal of Olympic Studies, v. 6, p. 93-106, 2022.
- ALMEIDA, WILLIAM DOUGLAS DE ; Veloso, R. C. ; RUBIO, K. . Olimpismo e memória: dos baús pessoais de atletas a um acervo público. PENSAR A PRÁTICA (ONLINE), v. 24, p. 1-16, 2021.
- RUBIO, K.; TABARINI, L. ; COSTA, M. R. L. . Quando a dor não faz parte do uniforme: os necessários cuidados com atletas e cuidadores no processo de reabilitação. REVISTA BRASILEIRA DE PSICOLOGIA DO ESPORTE, v. 8, p. 30-41, 2018.
- RABELO, IVAN ; RUBIO, KATIA . Literatura científica sobre mineração de textos aplicada à identificação da personalidade de atletas. Olimpianos - Journal of Olympic Studies, v. 2, p. 274-303, 2018.

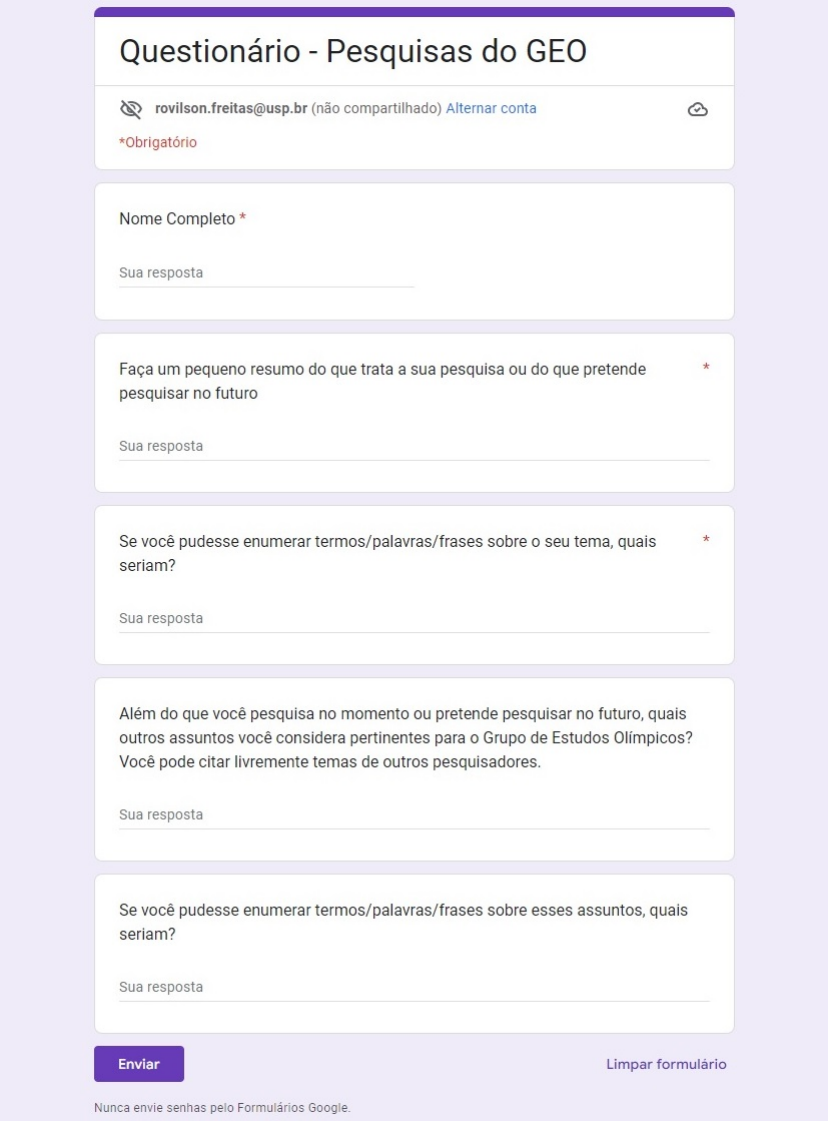
4.3.2 Entrevista não-estruturada com os especialistas

Foi realizada uma entrevista com os integrantes do GEO. Essa entrevista foi realizada através de um questionário online (Figura 7), com cinco perguntas abertas. É importante ressaltar que, dos que responderam o questionário, alguns integrantes estão em processo de pesquisa, outros ainda não tem tema definido, alguns já concluíram o processo e estão em vias de continuar o que já estudavam.

As perguntas oferecidas foram:

- Nome Completo
- Faça um pequeno resumo do que trata a sua pesquisa ou do que pretende pesquisar no futuro
- Se você pudesse enumerar termos/palavras/frases sobre o seu tema, quais seriam?
- Além do que você pesquisa no momento ou pretende pesquisar no futuro, quais outros assuntos você considera pertinentes para o Grupo de Estudos Olímpicos? Você pode citar livremente temas de outros pesquisadores.
- Se você pudesse enumerar termos/palavras/frases sobre esses assuntos, quais seriam?

Figura 7 – Formulário para pesquisa com os especialista do domínio



Fonte: Elaborado pelo autor

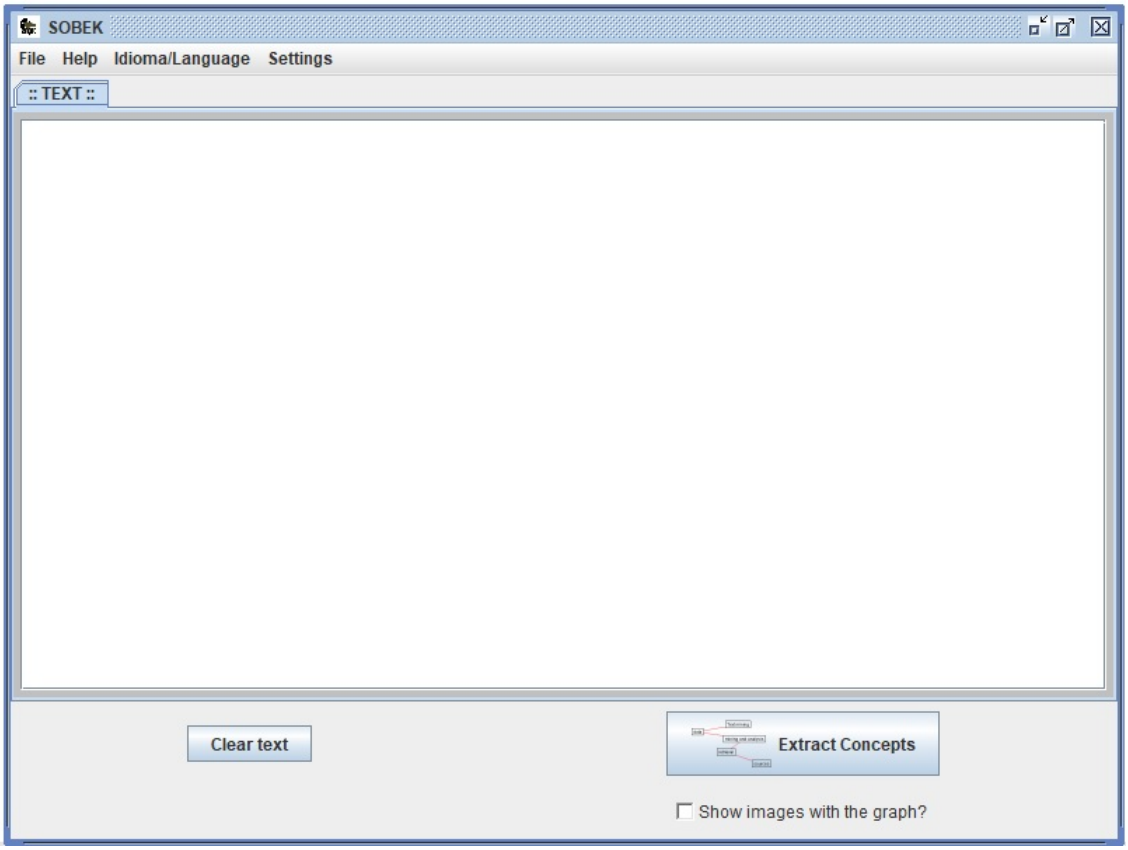
Foram 24 respostas, aproximadamente 80% dos integrantes atuais do GEO responderam ao questionário.

4.3.3 Extração de conhecimento via ferramenta semi-automática

Para esse trabalho, foi utilizada da ferramenta Sobek [1], desenvolvida na Universidade Federal do Rio Grande do Sul (UFRGS). Essa ferramenta (Figura 8) é capaz de identificar os conceitos relevantes em um texto a partir da análise de frequência desses termos no material textual. Embora a frequência não seja o único critério a considerar, pode ser importante indicativo de que determinado assunto possa ser relevante naquele contexto.

[1] http://sobek.ufrgs.br/

Figura 8 – Ferramenta SOBEK

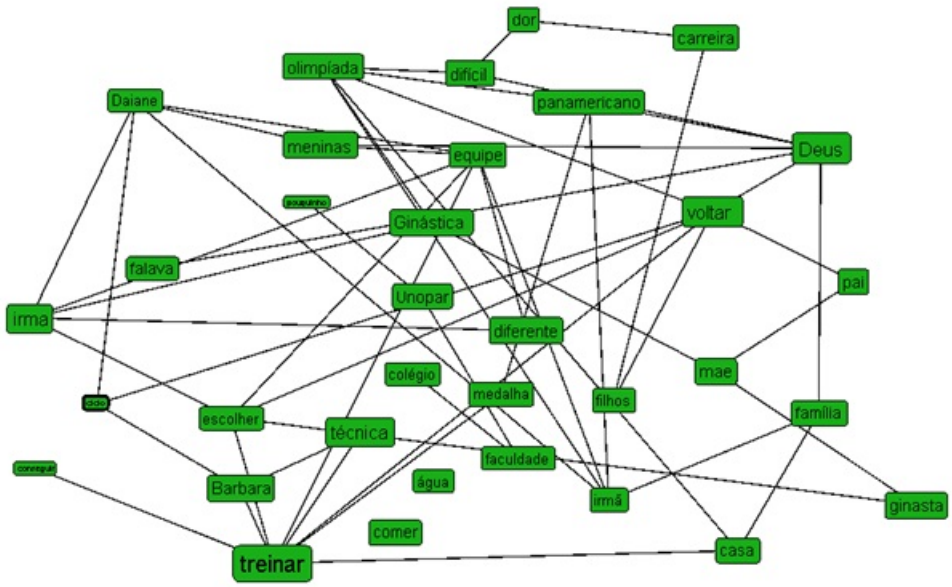


Fonte: Elaborado pelo autor

A ferramenta proporciona a seleção de um arquivo com uma lista de palavras que não devem ser consideradas), a configuração de quantos termos devem ser considerados e até mesmo a frequência mínima a ser verificada. Na figura 9, temos um exemplo de extração do conhecimento de uma atleta.

Alguns termos destacadas pela Sobek nesse exemplo, ajudam a consolidar temas que já foram tratados anteriormente pelo grupo. Nota-se o surgimento de assuntos como dor, família, formação acadêmica, entre outros. Importante reforçar que essa ferramenta usa como critério a frequência das palavras como fator de destaque. Esse é apenas um indicativo de que aquele assunto é relevante. Continua sendo fundamental o apoio dos especialistas das área que confirmam o que as ferramentas computacionais apresentam de resultado.

Figura 9 – Grafo gerado pela ferramenta Sobek



Fonte: Elaborado pelo autor

Uma vez que todo o material é analisado, os termos relevantes são selecionados. Após o processo de análise, é necessário separar os termos em três categorias: conceitos, verbos e relações. Em cada uma dessas categorias, será necessário criar um glossário, que será utilizado no próximo passo do processo.

As figuras 10, 11 e 12 trazem uma amostra desses glossários:

Figura 10 – Exemplo de Glossário de Conceitos

ID	Conceito
1	Pessoa
2	Atleta
3	Professor
4	Treinador
5	Saúde
6	Mental
7	Lesão
8	Dor
9	Raça
10	Gênero
11	Mulheres

Fonte: Elaborado pelo autor

Figura 11 – Exemplo de Glossário de Verbos

ID	Verbos
1	É_um(a)
2	Adquire_um
3	Cita_um
4	Controla_o
5	Está_em
6	Faz_o
7	Foi_em
8	Localizado_em
9	Sente_uma
10	Tem_uma
11	Treina_no
12	Estuda no
13	Trabalha Em

Fonte: Elaborado pelo autor

Figura 12 – Exemplo de Glossário de Relações

ID	Termo	Verbo	Termo
1	Atleta	É uma	Pessoa
2	Família	É uma	Pessoa
3	Atleta	Sente Uma	Dor
4	Atleta	Tem Uma	Lesão
5	Atleta	Sofre com	Racismo
6	Mulheres	Sofre com	Preconceito
7	Cidade	Localizada Em	Pais
8	Atleta	Treina no	Clube

Fonte: Elaborado pelo autor

4.4 Etapa 03 – Conceitualização

Segundo a etapa 3 da Ontoforinfoscience, deverá a partir dos glossários gerados na etapa anterior gerar os dicionários de conceitos e valores. Esse dicionário visa explicar os termos e verbos a partir de uma fonte da etapa de aquisição do conhecimento. Para a OntOlympic, foi seguido o padrão sugerido pela metodologia.

4.4.1 Dicionário de Conceitos

Como exemplo desse dicionário de conceitos, são apresentados alguns dos elementos importantes para a OntOlympic (Figura 13). Foram escolhidos os temas que já foram tratados por trabalhos anteriores do GEO, e estão disponíveis de forma completa no apêndice A:

Figura 13 – Exemplo de Dicionário de Conceitos

	Conceito	Sinônimos	Definição
1	Pessoa		Ser humano; quem pertence à espécie humana; criatura.
2	Atleta	Desportista	Pessoa treinada para competir, profissionalmente ou como amador, em exercícios, esportes ou jogos que requerem força, agilidade e resistência; esportista.
3	Professor		Aquele que leciona em algum estabelecimento de ensino; docente, mestre
4	Treinador	Técnico	Que ou aquele que treina ou dirige um time esportivo; técnico
5	Saúde		Estado do organismo com funções fisiológicas regulares e com características estruturais normais e estáveis, levando-se em consideração a forma de vida e a fase do ciclo vital de cada ser ou indivíduo. Bem-estar físico, psíquico e social.
6	Mental		estado de equilíbrio mental de um indivíduo, adaptado ao seu meio social e bem tolerante às condições e desafios da existência social e individual.
7	Lesão		Ato ou efeito de lesar. Ferimento ou traumatismo.

Fonte: Elaborado pelo autor

4.4.2 Tabela de Conceitos e Valores

Mendonça (2015) define a criação da tabela de conceitos e valores como o momento em que determinamos valores possíveis para cada um dos conceitos contidos no Dicionário de Conceitos (Figura 14), que também está na sua versão completa no apêndice A:

Figura 14 – Exemplo de Dicionário de Valores

	Conceito	Sinônimos	Definição	Valores
1	Pessoa		Ser humano; quem pertence à espécie humana; criatura.	Atleta, Família, Professor, Treinador
2	Atleta	Desportista	Pessoa treinada para competir, profissionalmente ou como amador, em exercícios, esportes ou jogos que requerem força, agilidade e resistência; esportista.	
3	Professor		Aquele que leciona em algum estabelecimento de ensino; docente, mestre	
4	Treinador	Técnico	Que ou aquele que treina ou dirige um time esportivo; técnico	
5	Saúde		Estado do organismo com funções fisiológicas regulares e com características estruturais normais e estáveis, levando-se em consideração a forma de vida e a fase do ciclo vital de cada ser ou indivíduo. Bem-estar físico, psíquico e social.	Mental, Lesões, Dor, Peso
6	Mental		estado de equilíbrio mental de um indivíduo, adaptado ao seu meio social e bem tolerante às condições e desafios da existência social e individual.	Sofre, Sofremos, Sofrimentos, Suicídio Traumatizada, Travada, Treme, Tremem, Tremendo, Depressão, Ansiedade

Fonte: Elaborado pelo autor

4.4.3 Dicionário de Verbos

A partir do glossário de verbos, definidos na etapa anterior, deve-se elaborar um dicionário com esses verbos, associando a sua aplicação do contexto da ontologia (Figura 15):

Figura 15 – Exemplo de Dicionário de verbos

ID	Verbo	Sinônimo	Exemplo de Uso
1	É_um(a)	Corresponde	O Atleta É_Uma pessoa
2	Adquire_um		
3	Cita_um	Fala	O atleta Cita_um Professor
4	Controla_o	Cuida	O Atleta Controla_O peso
5	Está_em		São Paulo Está_Em Região Sudeste
6	Localizado_em		São Paulo Localizado_Em Brasil
7	Sente_uma		O Atleta Sente_uma Dor
8	Tem_uma		O atleta Tem_Uma Profissão
9	Treina_no		O atleta Treina_no Clube

Fonte: Elaborado pelo autor

4.5 Etapa 04 - Fundamentação ontológica

Considerando a ontologia OntOlympic, essa provavelmente foi a fase da metodologia que demandou mais tempo e estudo por parte do desenvolvedor. Nessa etapa, seria necessário buscar ontologias já estabelecidas que serviram de fundamentação para a ontologia criada. Por se tratar de uma ontologia de domínio e de classificação leve, esse passo passa a ser opcional, segundo Santos (2016) Apud Mendonça (2015).

O reuso da ontologia de fundamentação deve ir além da sua importação em um editor de ontologias, é preciso manter o comprometimento ontológico com os princípios de fundamentação Mendonça (2015). Mesmo sendo um processo opcional para esse caso, foram buscadas alternativas para atender esse requisito. Por se tratar de um universo muito específico, não foram encontradas ontologias que atendessem e cumprissem com o compromisso ontológico dessa situação. Optou-se, portanto, pela não utilização de ontologia de fundamentação.

4.6 Etapa 05 - Formalização da ontologia

Um dos objetivos da criação de uma ontologia, é permitir que suas definições sejam interpretadas não apenas pelo humano, mas também por computadores. Com a ontologia definida, o computador será capaz de deduzir automaticamente uma informação a partir daquilo que foi determinado. Nessa etapa, portanto, temos a descrição formal da ontologia, saindo do conceitual para o ontológico-formal.

Santos (2016) propõe os seguintes passos para essa etapa do processo, que será realizada na ferramenta Protegé[2]:

a) Definir classes da ontologia: nesse passo são definidas as classes gerais da ontologia.

Na OntOlympic, foram definidas as seguintes classes principais (Figura 16):

Figura 16 – Classes principais da ontologia OntOlympic

Fonte: Elaborado pelo autor

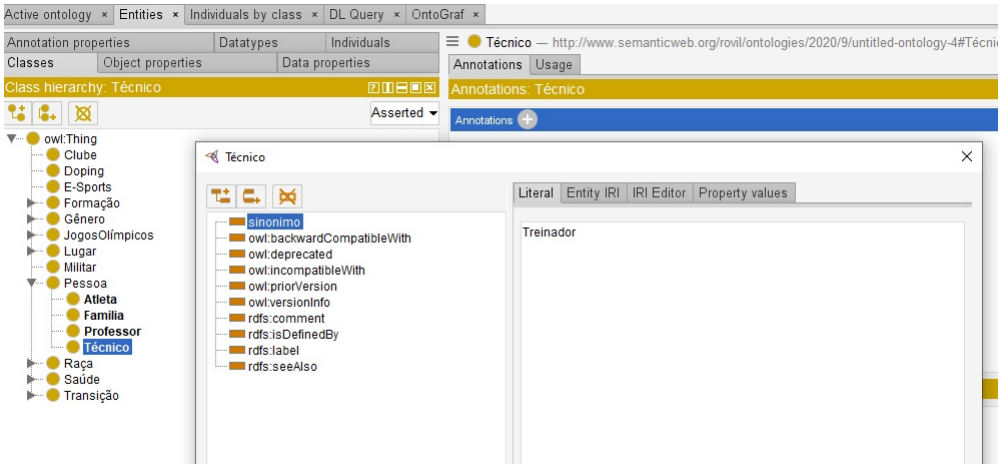
b) Introduzir os conceitos nas classes: nesse passo os conceitos levantados na etapa de aquisição das informações e estruturados no glossário de conceitos são distribuídos nas classes que foram definidas. Aqui podem ser definidas subclasses para a organização dos conceitos em uma taxonomia geral da ontologia.

c) Introduzir os sinônimos: o levantamento dos termos sinônimos foi feito no Dicionário de Conceitos desenvolvido na etapa de conceitualização deste trabalho.

Na figura 17, é mostrado o exemplo da inclusão do sinônimo da classe "Técnico", pertencente a classe pessoa. Como exemplo de sinônimo para essa classe, sugere-se o termo "Treinador".

[2] https://protege.stanford.edu/

Figura 17 – Exemplo de inclusão de sinônimo da classe Técnico

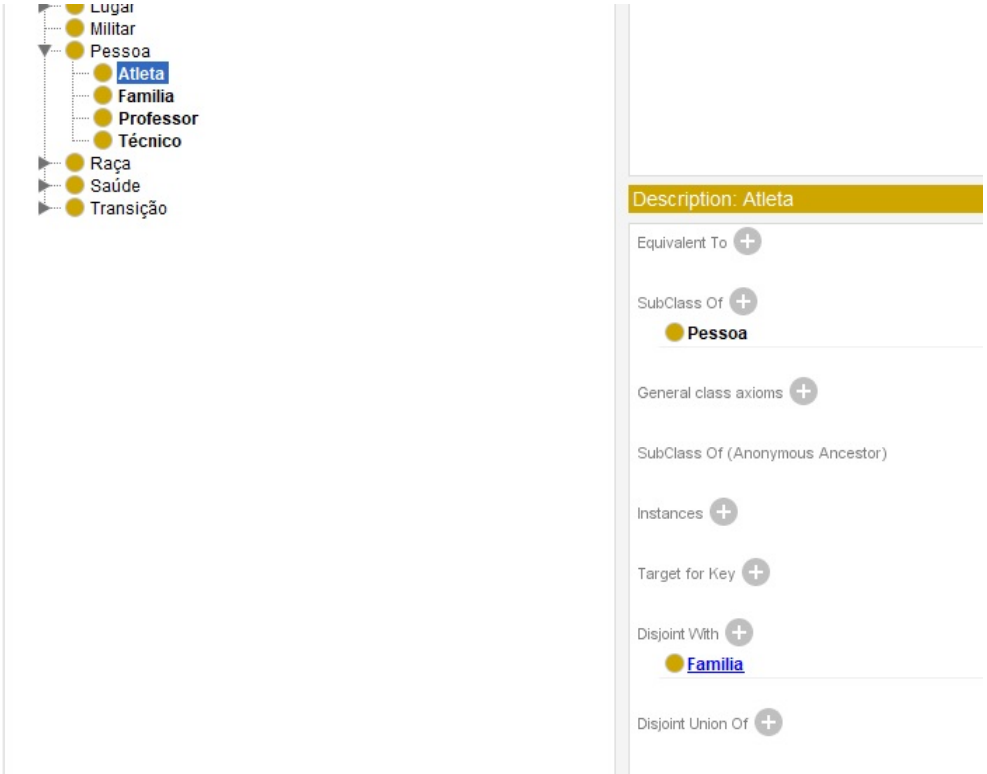


Fonte: Elaborado pelo autor

d) Definir as classes disjuntas: as classes disjuntas são aquelas que não possuem indivíduos em comum.

Na figura 18, como exemplo surge o caso da classes Atleta e Família. Pela lógica do domínio, o atleta não pode ser, ao mesmo tempo, seu próprio familiar. Portanto, essas classes precisam ser colocadas em disjunção, para que não ocorra problema nas inferências.

Figura 18 – Exemplo de disjunção entre classes Atleta e Família

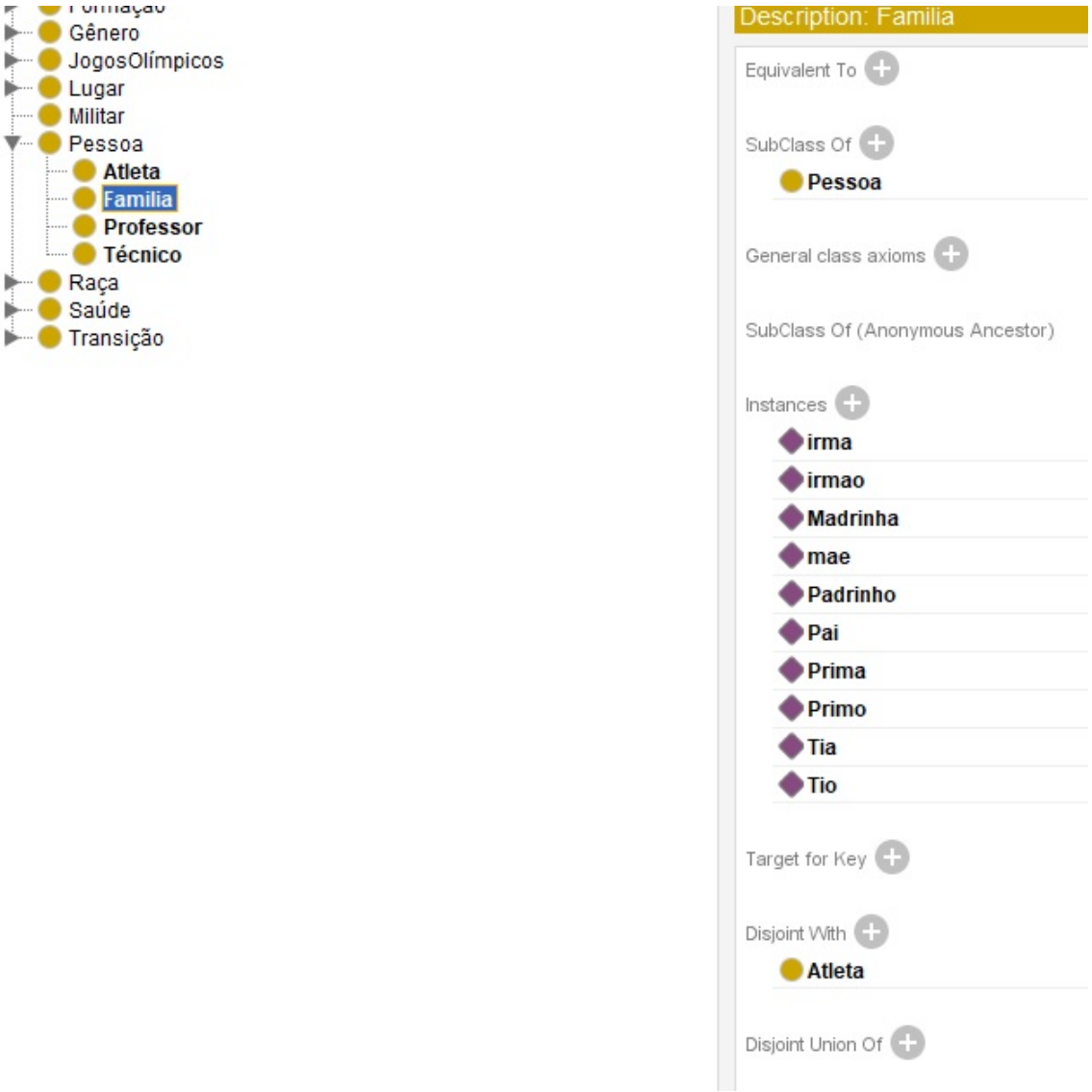


Fonte: Elaborado pelo autor

e) Criar instâncias: a definição de instâncias geralmente é um dos últimos passos a ser desenvolvido na ontologia, elas não são obrigatórias, e conforme a sua adesão irão definir o quão especifica é a ontologia. A Tabela de Conceitos e Valores desenvolvida na etapa de conceitualização deste trabalho deve ser utilizada como ponto de partida para identificar que conceitos possuem instâncias.

Na figura 19 o exemplo das instâncias da classe Família:

Figura 19 – Exemplo de instâncias da classe Família



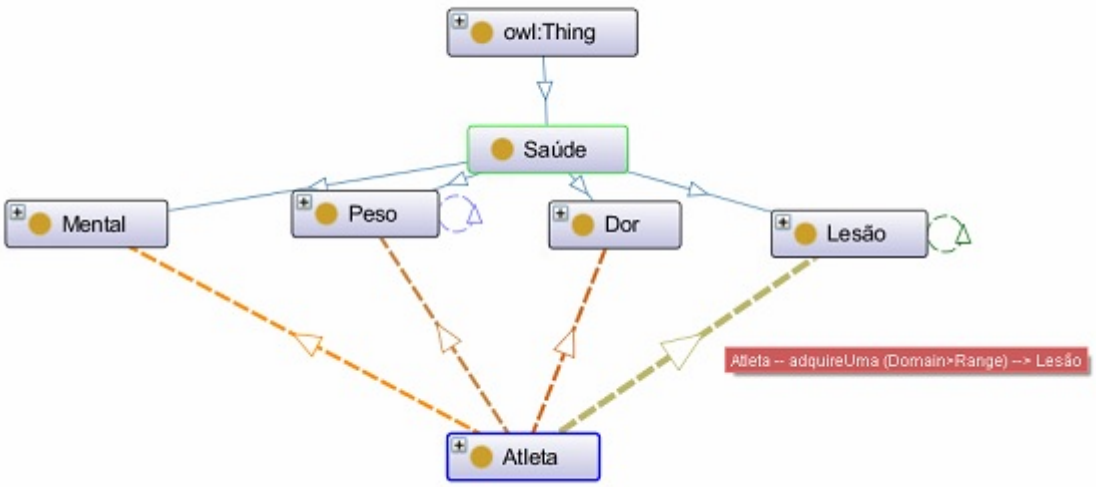
Fonte: Elaborado pelo autor

f) Definir propriedades de relação das classes: nesse passo são identificados, de forma mais aproximada, as relações das classes do domínio, conforme as ligações apresentadas no Glossário de Relações. Essas propriedades são nomeadas como propriedades de objeto ou dados.

A figura 20 mostra um exemplo de uma das relações entre classes propostas para a

ontologia OntOlympic, especificamente da classe Saúde. Nessa figura, existem relações entre Atleta com as subclasses Mental, Peso, Dor e Lesão. Uma das relações possíveis é Atleta -> AdquireUma -> Lesão.

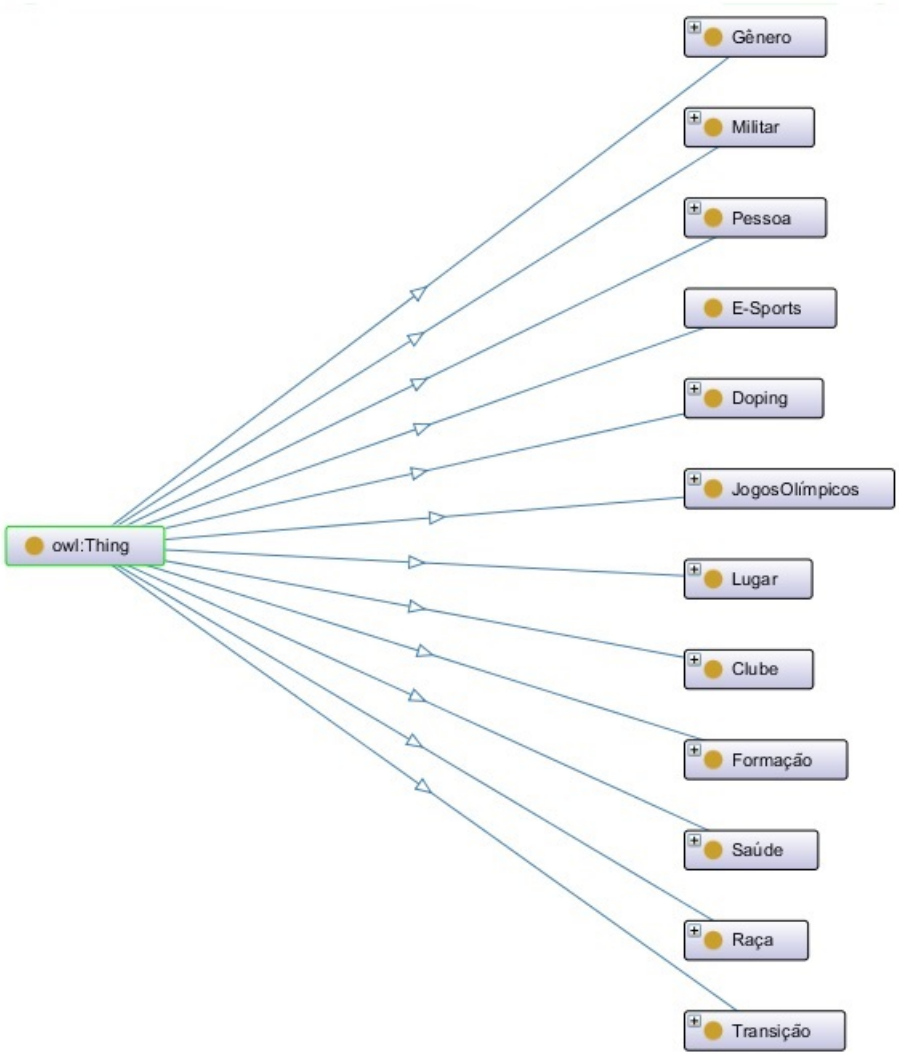
Figura 20 – Relações entre a classe Atleta e as classes relacionadas a Saúde



Fonte: Elaborado pelo autor

Abaixo, o esquema gráfico geral (figura 21) da OntOlympic. Em virtude de limitações de espaço, não foi possível representar a completude da ontologia, com todas as suas subclasses, relações e instâncias.

Figura 21 – Representação gráfica da OntOlympic

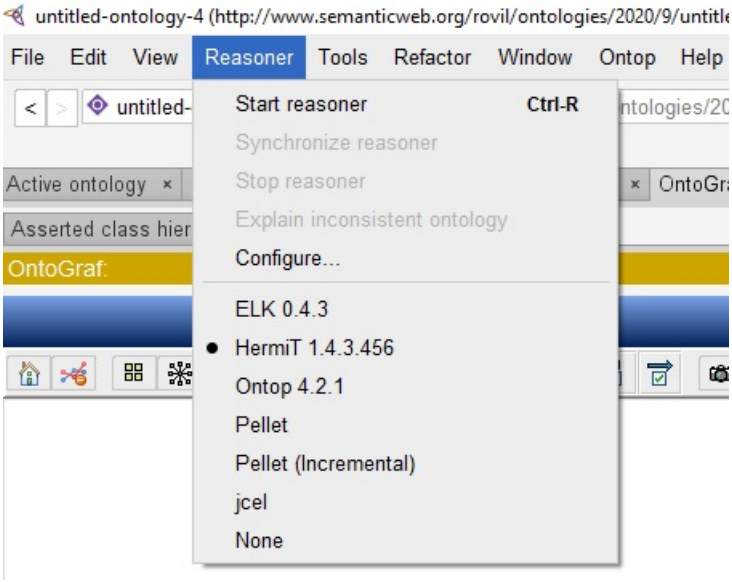


Fonte: Elaborado pelo autor

4.7 Etapa 06: Avaliação da Ontologia

A metodologia Ontoforinfoscience sugere uma série de elementos que podem ser utilizados como estratégia de avaliação. Basicamente, são dois passos a seguir para esse processo: no primeiro (validação), a avaliação serve para saber se a ontologia atende a realidade desse domínio. Se não existe elemento que não representa o domínio em questão. Uma vez estabelecido, o próximo passo é a validação do que foi executado (verificação). Isso significa verificar se todas as definições conceituais do domínio, estão de acordo com a lógica ontológica. Para a ontologia OntOlympic, uma avaliação preliminar foi realizada diretamente com os especialistas do Grupo de Estudos Olímpicos para a primeira fase (validação). Por se tratar de uma ontologia de domínio muito específico, algumas definições são praticamente exclusivas desse público. Já na fase da verificação, a sugestão é a utilização de classificadores automáticos do Protégé (Reasoner). Ele é capaz de mostrar os problemas que por ventura possam ter acontecido durante o processo de elaboração formal ontológico (Figura 22).

Figura 22 – Classificador (Reasoner da ferramenta Protégé)



Fonte: Elaborado pelo autor

4.8 Etapa 07: Documentação da Ontologia

Toda documentação correspondente à ontologia OntOlympic foi desenvolvida ao longo do processo. Mendonça (2015) sugere um modelo de tudo o que pode ser considerado para a documentação (Figura 23).

Figura 23 – Template norteadora para documentação da Ontologia

Etapa 1: Documento de Especificação
- Domínio/Escopo da ontologia - Propósito geral - Classes de usuários - Uso pretendido - Tipo da ontologia - Grau de formalidade - Questões de competência
Etapa 2: Documentos de referência
- Relação dos documentos do domínio tratado utilizados como materiais de referência para estudo e conceitualização do domínio.
Etapa 3: Modelos conceituais
- Conjunto de modelos conceituais desenvolvidos do domínio, que tenham sido aceitos e compartilhados pelos grupos envolvidos nesta etapa.
Etapa 4: Ontologias reutilizadas
- Relação e breve descrição das ontologias reutilizadas no desenvolvimento, que inclua a(s) ontologia(s) de fundamentação usada(s) como ponto de partida na construção.
Etapa 5: Conteúdo ontológico
- Taxonomia geral da ontologia - Dicionário de classes, que inclua como elementos: as definições textual e formal de cada classe, suas propriedades descritivas e lógicas, e referências às classes importadas de outras ontologias. - Dicionário de relações ontológicas, que inclua como elementos: as definições textuais e semiformal de cada relação, suas propriedades descritivas e lógicas, e referências às relações importadas de outras ontologias. - Estruturas gráficas de representação da ontologia, que apresentem em modo gráfico (visual) os relacionamentos existentes entre as classes da ontologia, tais como taxonomias, partonomias e outras estruturas.
Etapa 6: Métricas de avaliação
- Conjunto de **critérios, métodos e técnicas** usados na avaliação (validação e verificação) do conteúdo ontológico. - Conjunto de respostas às questões de competência propostas à ontologia.

Fonte: Elaborado pelo autor

4.9 Etapa 08: Disponibilização da ontologia em meio eletrônico

Segundo a metodologia Ontoforinfoscience, uma vez finalizada, a ontologia pode ser disponibilizada em algum ferramenta online para seus usuários. Essa ferramenta pode ser um site, por exemplo. Para esse trabalho, optou-se por disponibilizar a ontologia na ferramenta GitHub[3]. O link para acesso à ontologia é:

https://github.com/OntOlympic/OntOlympic

Futuramente, todos os documentos e arquivos referentes a esse trabalho estarão disponíveis no mesmo link.

4.10 Considerações Finais

Nesse capítulo, foi apresentado como a ontologia OntOlympic foi desenvolvida. A metodologia utilizada foi a Ontoforinfoscience, desenvolvido por Mendonça (2015). Nela são descritos todos os passos para a criação de uma ontologia, desde a análise da necessidade de elaboração, passando pela conceitualização e documentação, sua avaliação e disponibilização para a comunidade. O próximo capítulo trata da utilização dessa ontologia como suporte à anotação das entrevistas dos atletas olímpicos do GEO/USP.

[3] https://github.com/

CAPÍTULO

5

ANOTAÇÃO SEMÂNTICA BASEADA NA ONTOLOGIA ONTOLYMPIC E MINERAÇÃO DE TEXTOS

Para esse trabalho as entrevistas seriam submetidas aos algoritmos de mineração de texto: sem anotação semântica e depois, anotadas, com base na ontologia OntOlympic.

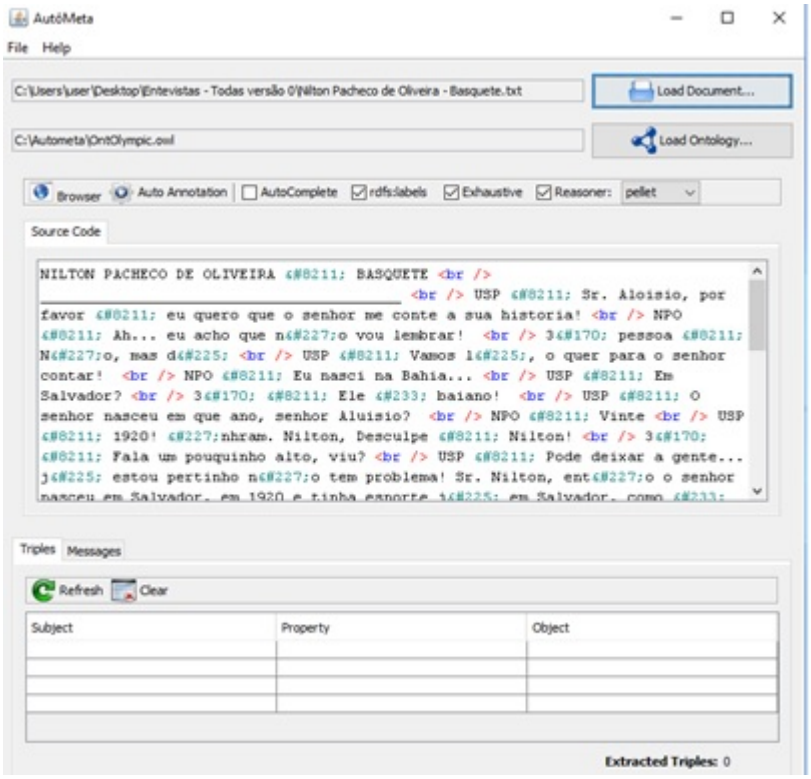
5.1 Anotação semântica com AutôMeta

No processo de pesquisa, foram testadas algumas alternativas para anotação semântica. Duas ferramentas em especial chamaram a atenção: a primeira foi a ferramenta GATE[1]. Essa ferramenta foi desenvolvida na Universidade de Sheffield e segundo Pereira (2014), permite ao desenvolvedor criar seus próprios recursos ou estender os que já existem. A ferramenta possuía, a priori, suporte para ontologias como base para anotação semântica. Porém, os *plugins* correspondentes a essa funcionalidade estavam descontinuados, segundo o próprio suporte da ferramenta.

A segunda ferramenta testada, foi a AutôMeta [2]. Desenvolvida por Fontes (2011), acabou se mostrando como a de melhor funcionamento para esse experimento em comparação ao GATE. A ferramenta AutôMeta conseguiu realizar as anotações nas entrevistas, bastando escolher a ontologia (nesse caso a OntOlympic). A AutôMeta possui duas interfaces: a primeira, visual (figura 24), onde é possível escolher a ontologia:

[1] https://gate.ac.uk/
[2] https://github.com/celsowm/AutoMeta

Figura 24 – Interface gráfica do AutôMeta



Fonte: Elaborado pelo autor

Também é possível ajustar de que maneira a anotação será realizada. Marcando a opção Exhaustive, todas as ocorrências no texto serão anotadas. Em caso de não marcação, apenas a primeira ocorrência do termo é anotado. A forma visual do programa não permite que mais de uma entrevista seja anotada. O trabalho deveria ser manual, entrevista por entrevista. Por isso, para esse trabalho, optou-se pelo AutôMeta via comandos (Figura 25), que permite que todas as entrevistas sejam anotadas automaticamente.

Figura 25 – Interface para o Autômeta via comandos

```
C:\Autometa\dist>java -jar Autometa.jar -ontology
"c:\Autometa\OntOlympic.owl"
-documentpath "c:\Autometa\2407\Todas"
-outpath "c:\Autometa\2407\Anotadas"
-exhaustive "true"
```

Fonte: Elaborado pelo autor

Nesse formato, é possível selecionar tanto a ontologia, a pasta de origem dos textos, quanto a pasta de destino dos textos anotados e a forma de anotação. Os nomes dos arquivos são

preservados, mas o tipo do arquivo é convertido para .htm. Com isso, foi necessário novamente converter os textos para o tipo .txt.

A AutôMeta anota as entrevistas no formato RDFa, que é um formato recomendado pela W3C para incorporar metadados ricos em uma página web.

Sobre as vantagens do padrão RDFa de anotação:

A vantagem da utilização do RDFa, segundo Silva *et al.* (2018), é que máquinas de buscas podem melhorar seus resultados aumentando a precisão sobre o real significado de um documento. Ou seja, as máquinas de buscas podem agregar os dados de um documento com dados de outro documento, enriquecendo os resultados de buscas.

O padrão RDFa é composto por 4 atributos: (Prefix, Resource, Property e Typeof). Prefix descreve os vocabulários reusados no HTML. Resource descreve os recursos. Property relaciona dois elementos e Typeof para representar o tipo de um elemento.

Na Figura 26, temos um exemplo de como fica a anotação gerada pelo AutôMeta em uma entrevista deste trabalho. Ali, temos o termo “Munique”. A palavra Munique tem uma representação importante no contexto das entrevistas: é uma cidade localizada na Alemanha, que fica na Europa. Também foi sede de uma edição dos Jogos Olímpicos no ano de 1972.

Figura 26 – Formato de anotação semântica RDFa

```
<span id='am-72' about='untitled-ontology-4:Munique'
    typeof='owl:Thing untitled-ontology-4:Sede untitled-ontology-4:Cidade
    untitled-ontology-4:JogosOlímpicos untitled-ontology-4:Lugar'>
    <span id='am-73' rel="untitled-ontology-4:foiSedeEm" resource="
    untitled-ontology-4:1972"></span>
    <span id='am-74' rel="untitled-ontology-4:localizadoEm" resource="
    untitled-ontology-4:Europa"></span>
    <span id='am-75' rel="untitled-ontology-4:localizadoEm" resource="
    untitled-ontology-4:Alemanha"></span>
    Munique
</span>
```

Fonte: Elaborado pelo autor

Essa anotação foi baseada nos parâmetros definidos para a ontologia OntOlympic. Dentro da ferramenta Protégé, o indivíduo “Munique” tem as seguintes informações: Pertencem às Classes Sede e Cidade, têm as relações de localização e em que ano a edição se realizou (Figura 27):

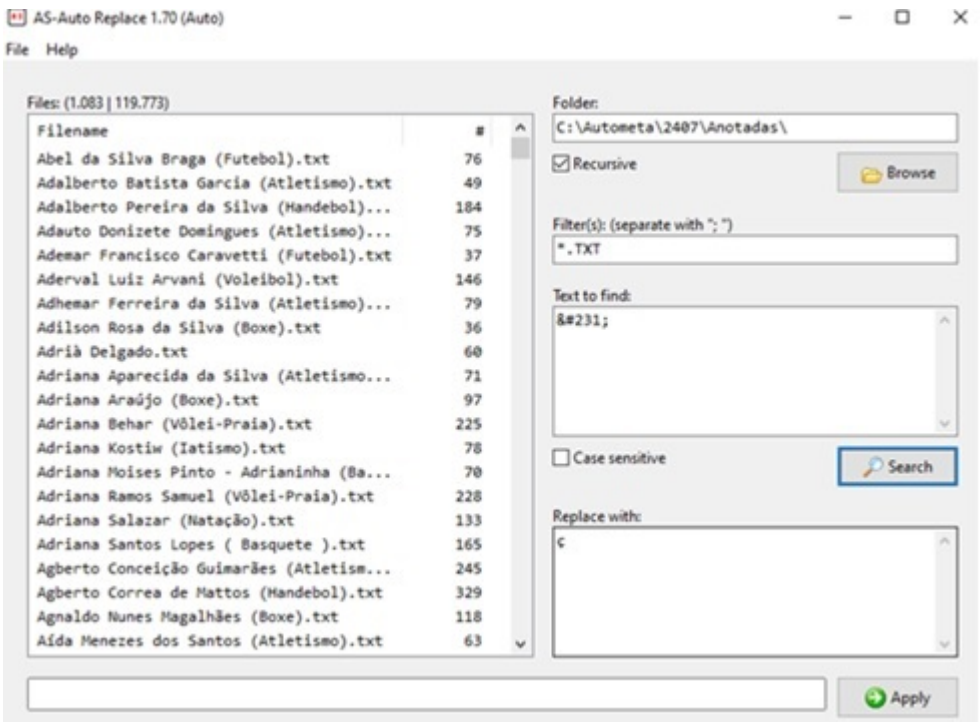
Figura 27 – O indivíduo "Munique"e suas classes e relações correspondentes

Fonte: Elaborado pelo autor

A relação de localização de “Europa”, veio automaticamente pela inferência, visto que Munique LocalizadoEm Alemanha, que está LocalizadaEm Europa. Portanto, Munique está LocalizadoEm Europa. Por último, foi necessário um ajuste nas entrevistas anotadas. Elas não reconheceram a acentuação e os símbolos em língua portuguesa. Cada uma delas foi substituída por um código. Para solucionar esse problema de forma rápida, foi utilizada a ferramenta AS Auto Replace. [3], programa que permite que vários arquivos tenham algum termo substituído de uma vez (Figura 28).

[3] https://www.andreas-software.com/international/program$_{a}s_{a}uto_{r}eplace.php$

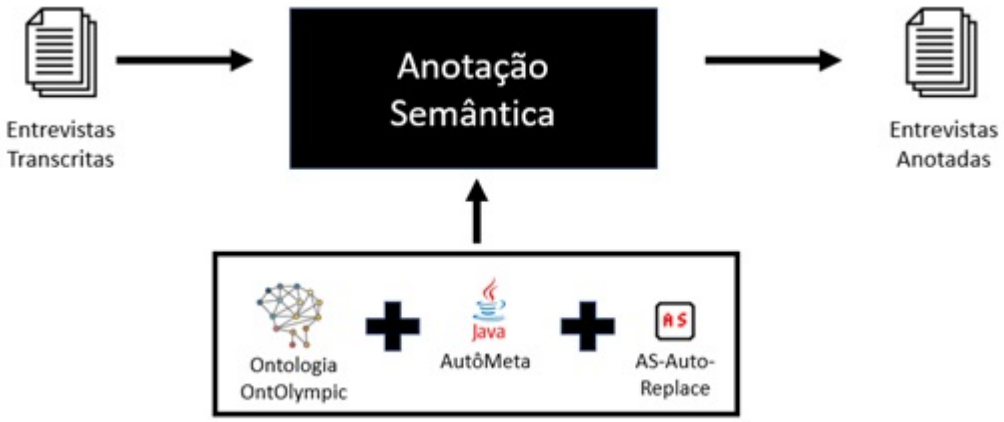
Figura 28 – Ferramenta AS Auto-Replace, usada para ajustes da acentuação



Fonte: Elaborado pelo autor

Com isso, temos o processo completo (Figura 29) de anotação semântica nas entrevistas do Grupo de Estudos Olímpicos baseada na ontologia OntOlympic. As entrevistas estão, portanto, preparadas para o processo de mineração de textos, realizada a seguir.

Figura 29 – Metodologia para anotação semântica

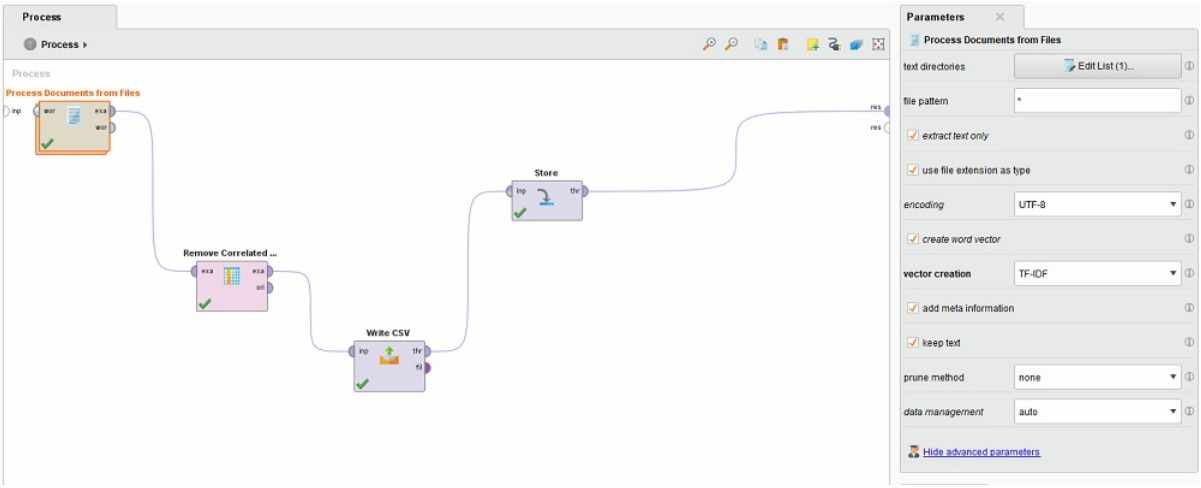


Fonte: Elaborado pelo autor

5.2 Mineração de textos com Rapidminer

Para os experimentos com mineração de textos, foi utilizada a ferramenta Rapidminer[4]. É um programa para mineração de dados que possui suporte a mineração de textos. Tem a vantagem de ser um programa essencialmente visual, não exigindo conhecimentos avançados de programação para quem o opera. Os elementos visuais são selecionados pelo usuário que apenas deve configurar os parâmetros correspondentes ao teste (Figura 30).

Figura 30 – Exemplo de estrutura no Rapidminer para criação de Matriz TF-IDF



Fonte: Elaborado pelo autor

Para o experimento envolvendo mineração de texto no Rapidminer, foram considerados os seguintes critérios:

- Criação de uma tabela documento-termo utilizando o índice TF-IDF. Nessa tabela, cada linha representa uma entrevista, e cada coluna, o respectivo termo.

  Nesse passo, temos o pré-processamento. Para esse trabalho, foram utilizados os seguintes elementos:

  - Tokenização: processo de individualização das palavras.
  - Aplicação de StopWords: uma lista de palavras não relevantes para o processo foi inserida, visando reduzir o tamanho da matriz.
  - Stemming: o processo de stemming reduz as palavras flexionadas a sua raiz.
  - Remover correlatas: Uma ferramenta do Rapidminer reduz palavras relacionadas entre si dentro da matriz. Isso faz com que termos parecidos não sejam repetidos.

[4] https://rapidminer.com/

Com todos os critérios ajustados, a matriz é gerada e armazenada num repositório.

- Utilização do algoritmo K-means para o agrupamento.
  - A matriz TF-IDF gerada na etapa anterior que está armazenada num repositório, é recuperada.
  - A ferramenta correspondente ao algoritmo de agrupamento é selecionada no Rapidminer. Nesse caso, é a ferramenta Clustering (K-Means).
  - Um dos pré-requisitos para utilização do K-means, é a inserção do número máximo de clusters (k) para análise. Para esse experimento, foi escolhido o número k=50. Esse número foi escolhido por ter sido o maior valor com processamento considerado possível para testes. Acima desse valor, o processamento acabou se tornando extremamente lento.
  - Com a escolha do k=50, teremos o processamento dos agrupamentos variando de 2 a 50. O Rapidminer oferece uma solução (Loop Parameters) que realiza esses testes de forma automática, gerando todos os resultados necessários.
  - Outro parâmetro importante a ser escolhido nesse passo é o Numerical Measure. Esse parâmetro determina o critério de similaridade no espaço vetorial criado na matriz TF-IDF. O critério mais utilizado normalmente é o Euclidiano. Porém, ele não se demonstra efetivo para a representação textual, podendo causar alguma distorção no resultado. Portanto, para esse experimento o critério de similaridade escolhido foi o Cosseno, que tem a tendência de representar melhor a similaridade para esse contexto, visto sua análise pelo ângulo.
  - A análise de performance dos agrupamentos utilizando o índice de Davies-Bouldin, também oferecido pelo Rapidminer. O índice é calculado para todos os testes de agrupamentos entre 2 e 50. Quanto menor o valor, teoricamente melhor é o agrupamento.

Como resultados, temos :

- Matriz com a relação Agrupamento x Entrevista (Figura 31). É considerada nesse caso o melhor agrupamento dado pelo índice Davies-Bouldin.

Figura 31 – Exemplo de Matriz Agrupamento x Entrevista

Cluster	Atleta
cluster_0	Atleta01 (tênis de quadra).txt
cluster_0	Atleta02 (Tenis).txt
cluster_0	Atleta03 (tênis de quadra).txt
cluster_0	Atleta04 (Tenis de mesa).txt
cluster_0	Atleta05 (tênis de quadra).txt
cluster_0	Atleta06 (tênis de quadra).txt
cluster_0	Atleta07 (tênis de quadra).txt
cluster_0	Atleta08 (tenis-quadra).txt
cluster_1	Atleta09(Tiro esportivo).txt
cluster_1	Atleta10(Tiro esportivo).txt
cluster_1	Atleta11(Tiro esportivo).txt
cluster_1	Atleta12(Tiro esportivo).txt
cluster_1	Atleta13(Tiro Esportivo).txt
cluster_1	Atleta14(tiro esportivo).txt
cluster_1	Atleta15(Tiro Esportivo).txt
cluster_1	Atleta16(Tiro esportivo).txt
cluster_1	Atleta17(Tiro Esportivo).txt

Fonte: Elaborado pelo autor

- Considerando o item anterior, a matriz de palavras de cada *cluster*, reunindo os possíveis termos que podem justificar a razão daquele agrupamento (Figura 32).

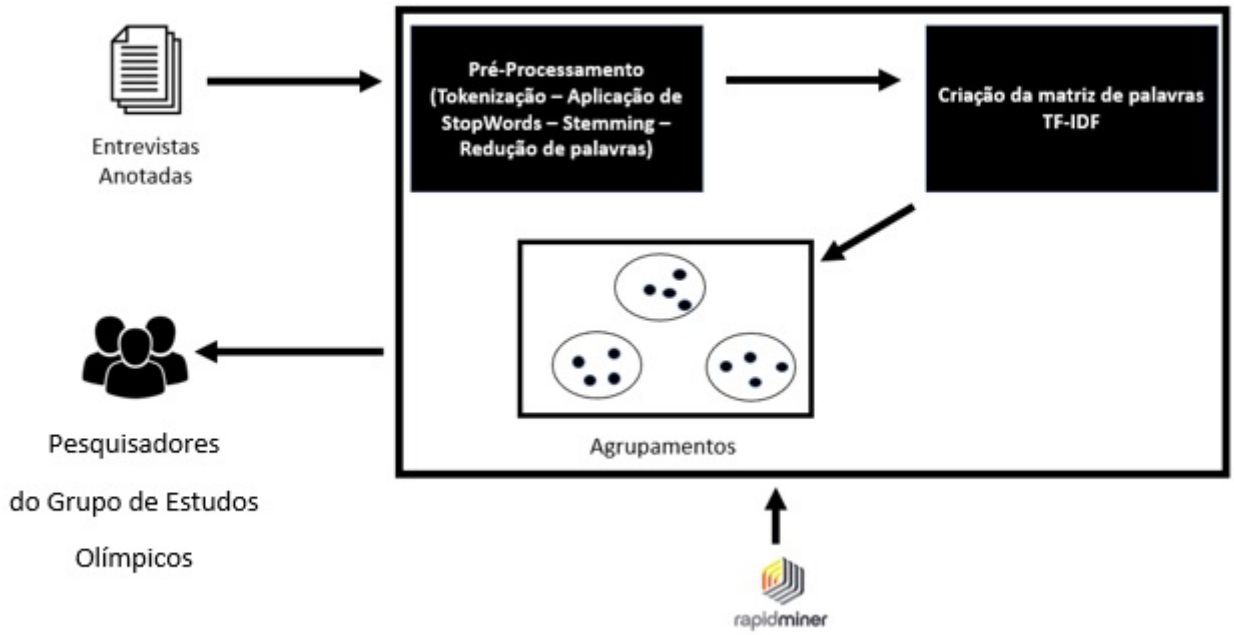
Figura 32 – Exemplo de Matriz Atributo x *Cluster* (Utilizando o índice Davies-Bouldin)

Attribute	cluster_0	cluster_1	cluster_2	clus... ↓	cluster_4	cluster_5	cluster_6	cluster_7	cluster_8
futebol	0.018	0.006	0.009	0.151	0.037	0.001	0.003	0.012	0.016
corinthia...	0.001	0.001	0.008	0.108	0.014	0	0	0.006	0.002
guaran	0.001	0.001	0.003	0.082	0.002	0	0	0.002	0
jogador	0.025	0.001	0.003	0.081	0.044	0	0.001	0.001	0.019
treinador	0.023	0.008	0.019	0.080	0.019	0.019	0.021	0.060	0.020
sant	0.012	0.011	0.012	0.074	0.010	0.009	0.005	0.011	0.012
feminin	0.025	0.002	0.009	0.073	0.011	0.002	0.009	0.011	0.024
jog	0.077	0.003	0.005	0.067	0.039	0.001	0.001	0.005	0.062
campin	0.020	0.009	0.007	0.063	0.003	0.002	0	0.016	0
palmeir	0.004	0.001	0.001	0.045	0.005	0	0	0	0
vasc	0.001	0.001	0.012	0.043	0.025	0.005	0	0	0.008
goleir	0.001	0.002	0.001	0.043	0.030	0	0.000	0.001	0.029
portugues	0	0	0.001	0.040	0.002	0	0	0.002	0

Fonte: Elaborado pelo autor

Com isso, concluí-se o processo de agrupamento, que pode ser representado pela figura 33:

Figura 33 – Metodologia para agrupamentos com documentos anotados



Fonte: Elaborado pelo autor

5.3 Considerações Finais

Foi apresentado nesse capítulo a utilização da ontologia OntOlympic na anotação semântica das entrevistas, utilizando a ferramenta AutôMeta. As entrevistas anotadas incorporaram as informações da ontologia. Depois da anotação, as entrevistas foram submetidas ao processo de mineração de textos no Rapidminer. O algoritmo escolhido para esses testes foi o de agrupamento (k-means). No próximo capítulo, foi realizado um teste experimental com as entrevistas anotadas agrupadas em comparação com as entrevistas sem anotação. Esse estudo visa perceber se existe alguma distinção nos agrupamentos encontrados em cada teste. A hipótese é de que a experiência envolvendo as entrevistas anotadas pode trazer grupos mais significativos do que os grupos das entrevistas não anotadas.

CAPÍTULO

6

ESTUDO EXPERIMENTAL

6.1 Teste sem anotação semântica

Foram utilizadas no experimento todas as entrevistas transcritas do acervo do GEO (1083 arquivos). Os primeiros testes foram feitos sem a anotação semântica.

Considerando as 1083 entrevistas, o melhor resultado do índice Davies-Bouldin é o que representa 47 *clusters*. Lembrando que para esse índice, quanto menor o valor, melhor o agrupamento é representado (Figura 34).

Figura 34 – Resultado do índice Davies-Bouldin para teste sem anotação semântica baseada em ontologia

Número de Clusters	Índice Davies-Bouldin
47	3,475
2	3,591
41	3,65
35	3,683
44	3,725
49	3,77
40	3,77
42	3,795
38	3,806
37	3,831

Fonte: Elaborado pelo autor

O número de *clusters* ideal nos dez primeiros resultados ficou, basicamente entre 38 e 49 agrupamentos. Apenas o número 2 destoou desse resultado.

Ao analisar os grupos gerados, nota-se claramente o fator modalidade como critério para o agrupamento. Possivelmente, por se tratar da mesma modalidade, atletas usam termos que são comuns a ela. Equipamentos esportivos, vestimentas, técnicos em comum, locais bem característicos são elementos que aproximam essas entrevistas. Por exemplo, o agrupamento 0, dos 47 grupos gerados. Ao fazer a leitura dos termos desse *cluster* (dado pelo resultado gerado pelo K-Means), temos as seguintes palavras como dominantes (Figura 35):

Figura 35 – Palavras dominantes do grupo 25

Atributo	Valor
biciclet	1,175
bik	0,171
mountain	0,124
jô	0,089
viços	0,089
pedal	0,071
ciclism	0,059
dai	0,052
arapong	0,051
calo	0,05
prov	0,048
pist	0,045
levant	0,043
andar	0,036

Fonte: Elaborado pelo autor

Temos aqui termos como: Bicicleta, Bike, Mountain, pedal, ciclismo, Caloi (uma marca conhecida de bicicletas). Percebemos também a presença de Viçosa e Levantamento. Viçosa (MG) é um centro importante para o levantamento de peso no Brasil. Ao verificar as entrevistas agrupadas para esse *cluster*, percebemos que o grupo gerado é dominado pelas modalidades Ciclismo e Levantamento de peso (Figura 36):

Figura 36 – Atletas que compõe o grupo 0

Cluster	Atleta
cluster_0	Atleta01 (Ciclismo).txt
cluster_0	Atleta02 (Ciclismo).txt
cluster_0	Atleta03 (Futebol).txt
cluster_0	Atleta04(Ciclismo).txt
cluster_0	Atleta05 (Ciclismo).txt
cluster_0	Atleta06 (Levantamento de peso).txt
cluster_0	Atleta07(Ciclismo).txt
cluster_0	Atleta08(Ciclismo).txt
cluster_0	Atleta09 (Ciclismo).txt
cluster_0	Atleta10 (Ciclismo) .txt
cluster_0	Atleta 11 (Levantamento de Peso) doc.txt

Fonte: Elaborado pelo autor

Portanto, os agrupamentos se justificam basicamente pelas modalidades nesse momento. Existe a hipótese de que, ao incluir a anotação semântica baseada em ontologia nas entrevistas, os resultados dos agrupamentos se tornem melhores, tanto no índice de avaliação (Davies-Bouldin), quanto na análise das entrevistas por especialistas.

6.2 Teste com anotação semântica

As entrevistas, agora anotadas, passaram pelo mesmo processo. A primeira coisa a se notar, em comparação ao teste sem anotação, é que o índice Davies-Bouldin tem ligeira queda. O número de *clusters* para esse caso também variou: de 47 para 44 grupos (Figura 37).

Figura 37 – Resultado do índice Davies-Bouldin para teste sem anotação semântica baseada em ontologia

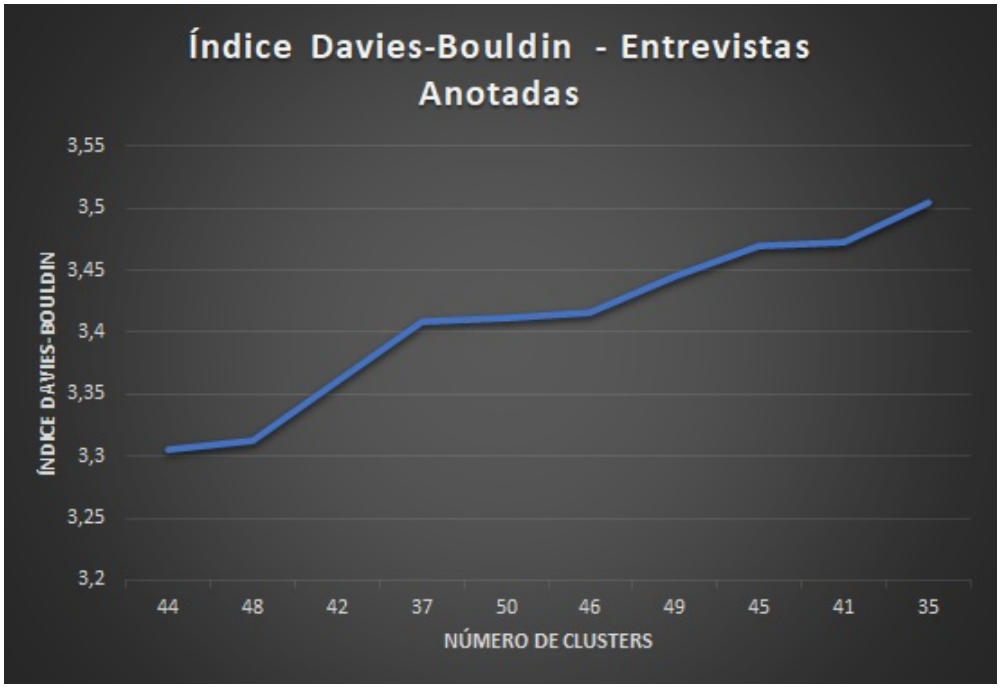
Número de Clusters	Índice Davies-Bouldin
44	3,305
48	3,313
42	3,361
37	3,409
50	3,412
46	3,416
49	3,445
45	3,47
41	3,473
35	3,504

Fonte: Elaborado pelo autor

Pode-se notar o fato de que os dez primeiros resultados estão mais estáveis, considerando o número de grupos. No teste com anotação, percebemos a variação entre 35 e 50, ao passo que no teste sem anotação, a faixa numérica é bem parecida, porém traz o número 2 fora da curva numérica que seria esperada.

Nota-se também que com a aplicação da anotação semântica, os valores correspondentes ao índice Davies-Bouldin variaram menos em relação aos valores do experimento sem anotação. Se no teste sem anotação, a variação foi entre 3,475 e 3,831 (0,356) mostrada na figura 38, no caso do teste com as entrevistas anotadas, a variação foi entre 3,305 e 3,504 (0,199), mostrada na figura 39. Isso pode ser um indicativo de que os agrupamentos com anotação semântica podem estar melhor organizados do que sem anotação semântica.

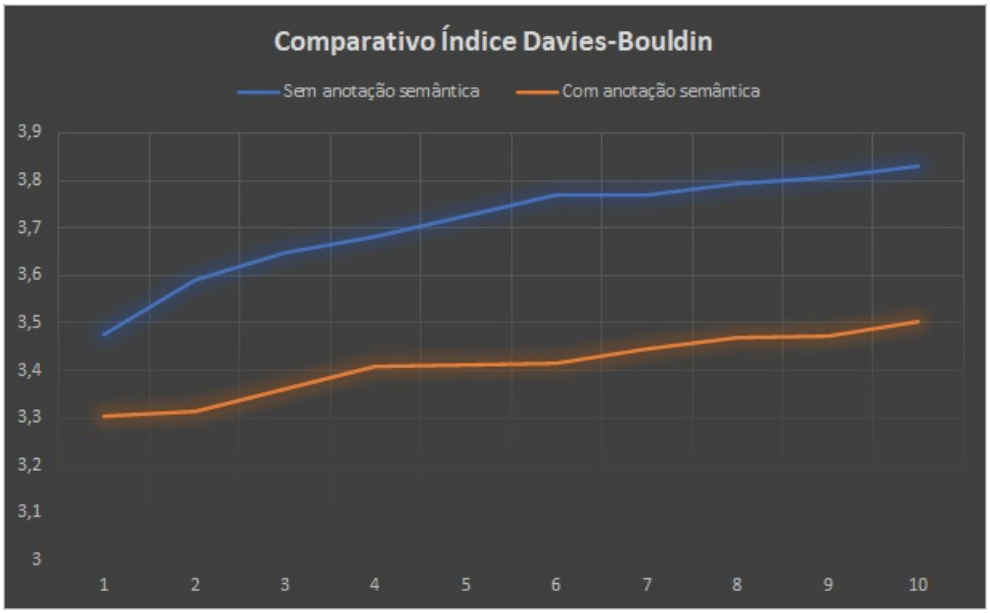
Figura 38 – Gráfico - Davies-Bouldin com anotação semântica



Fonte: Elaborado pelo autor

O gráfico abaixo (Figura 41), compara os dois resultados:

Figura 39 – Gráfico Comparativo - Davies-Bouldin



Fonte: Elaborado pelo autor

Considerando os agrupamentos, nota-se claramente que, embora o fator modalidade ainda seja bastante importante, com a anotação semântica os grupos estão menos exclusivos. Isso significa que o critério de agrupamento não é exclusivamente a modalidade, visto que os atletas agora se misturam.

Tomamos como exemplo o grupo 25 (Figura 40) do teste com anotação semântica (considerando que para esse teste, o número de *clusters* considerado ideal é 44).

Figura 40 – Atletas que compõe o grupo 25

Cluster	Atleta
cluster_25	Atleta01(Ginastica Artistica).txt
cluster_25	Atleta02 (Ginastica Artistica).txt
cluster_25	Atleta03 (Esgrima).txt
cluster_25	Atleta04 (Esgrima).txt
cluster_25	Atleta05 (Ginastica artistica).txt
cluster_25	Atleta06 (Ginástica artística).txt
cluster_25	Atleta07 (Futebol).txt
cluster_25	Atleta08 (Ginastica Artistica).txt
cluster_25	Atleta09(ginástica artística).txt
cluster_25	Atleta10 (Polo Aquático).txt
cluster_25	Atleta11(Natação).txt
cluster_25	Atleta12 (ginastica-artistica).txt
cluster_25	Atleta 13 (ginástica artística).txt
cluster_25	Atleta 14 (volei-praia).txt
cluster_25	Atleta15 (volei-quadra).txt
cluster_25	Atleta 16 (Esgrima).txt
cluster_25	Atleta17 (ginástica artística).txt

Fonte: Elaborado pelo autor

Primeiramente, nota-se uma diversidade maior de modalidades, o que significa que esse critério não foi o mais importante para esse caso com anotação semântica. Isso também ocorreu em outros grupos. Portanto, esse grupo parece ser mais interessante do ponto de vista de análise, pois pode representar algum conhecimento ainda não explorado pelo GEO, ou pode confirmar alguma hipótese previamente sugerida.

Ao analisar os termos dominantes do *cluster* (Figura 41), podemos chegar a alguns pontos importantes:

Figura 41 – Palavras dominantes do grupo 25

Grupo 25	
curitib	ginast
gin	treinador
daian	aten
stic	mulh
pequim	arbitr
jad	aleg
aparelh	fisioterap
luiz	sio
mestr	permanent
antoni	gargalh
joelh	espanh
cirurg	flameng
ombro	chin
trav	reg
oleg	
barcelonet	
menin	
estrutur	
faculdad	

Fonte: Elaborado pelo autor

Percebemos alguns termos importantes e que são associados à anotação semântica baseada na ontologia OntOlympic: percebe-se alguns termos relacionados a lugares (Curitiba, Pequim, Espanha), que podem indicar algum tipo de padrão ou deslocamento. Termos relacionados a questões de saúde (joelho, ombro, cirurgia, fisioterapia), que podem refletir questões ligadar a lesões. Citação a treinadores, formação acadêmica e questões de gênero. Os pesquisadores, portanto, podem se aprofundar nessas entrevistas visando atender esses temas em especial ou relacioná-los com outros em potencial.

6.3 Considerações Finais

O estudo experimental, apresentado nesse capítulo, demonstrou que os resultados dos testes com mineração de texto das entrevistas anotadas trouxeram ligeira melhora em comparação ao mesmo teste realizado com entrevistas sem anotação. Embora tenha sido considerado um índice numérico de qualidade dos agrupamentos (Davies-Bouldin) como um dos critérios, é importante levar em consideração a análise dos grupos resultantes do teste. Nota-se uma qualidade maior dos agrupamentos, que sai de um padrão óbvio de apenas agrupar por modalidade, para agrupamentos que envolvem temas mais específicos, como os relacionados a lesões, lugares, peso corporal, entre outros. Esse resultado pode contribuir para um direcionamento mais objetivo da análise dos pesquisadores do GEO/USP.

CAPÍTULO

7

CONSIDERAÇÕES FINAIS

Analisar entrevistas não-estruturados traz grandes desafios para os pesquisadores das mais diversas áreas do conhecimento. Uma solução computacional pode resultar na economia de tempo, recurso tão caro a quem realiza investigações acadêmicas. Além disso, pode proporcionar resultados mais relevantes, visto que no trabalho manual, alguns detalhes podem ser ignorados.Um suporte automatizado pode trazer a luz algo que não foi percebido de forma clara.

O presente trabalho explora o uso de anotação semântica baseada em ontologia para enriquecimento de textos extraídos de entrevistas não-estruturadas e análise por meio de técnicas de agrupamento de textos. A solução proposta inclui a nova ontologia de domínio OntOlympic, criada para contemplar aspectos de interesse para pesquisas relacionadas a atletas olímpicos brasileiros, especialmente do Grupo de Estudos Olímpicos da Universidade de São Paulo. A metodologia aplicada para anotação semântica, pré-processamento dos textos, agrupamento e avaliação.

O estudo experimental indica uma tendência de melhora nos resultados de tarefas de agrupamento quando aplicadas a entrevistas anotadas. Mais do que o resultado apenas matemático, foi percebido uma melhora na composição dos grupos que tiveram o processo de anotação. Muda-se a tendência de um critério básico (no caso desse experimento, da separação das entrevistas apenas por modalidade), e passa a trazer grupos mais diversos e com mais possibilidades de novos conhecimentos.

Esse resultado reflete o conhecimento já adquirido em trabalhos concluídos e em andamento desenvolvidos por pesquisadores do grupo de pesquisa que disponibilizou o acervo de entrevistas.

Com esse trabalho, surge a oportunidade de adaptação para outros grupos de estudos que tem características próximas ao GEO. Pesquisadores que trabalham com textos ou entrevistas não-estruturadas, podem adaptar a abordagem proposta aqui, considerando suas particularidades e obviamente, seu domínio, que precisaria ser avaliado.

7.1 Trabalhos futuros

Como perspectiva para trabalhos que venham a ser desenvolvidos no futuro, podemos considerar uma plataforma para a formalização da ontologia. O processo de desenvolvimento implica numa série de passos, que envolvem a criação de tabelas, associações, etc. Isso poderia ser feito usando uma ferramenta que facilitasse esse processo. Isso evitaria trabalhos desnecessários e repetitivos no processo. Já existem alguns trabalhos acontecendo nesse momento, mas poderia ser uma evolução no processo de criação ontológica.

Uma possibilidade de trabalho futuro, seria desenvolver uma aplicação voltada ao usuário final para facilitar o acesso às informações geradas pelo processo proposto nesse trabalho.

Outro potencial trabalho para o futuro poderia ser a ampliação dos testes em mineração de textos. Como hipótese, outras estratégias poderiam ser aplicadas (como classificação, agrupamentos hierárquicos, entre outros), além da aplicação de outros índices de avaliação de qualidade dos agrupamentos (como silhouette, dunn, entre outros). Além é claro, da adaptação do processo dessa trabalho para um outro domínio, que traria mais dados para futuras análises.

REFERÊNCIAS

ARRUDA, C. G. de. **SOM4SImD : um método semântico baseado em ontologia para detectar similaridade entre documentos**. 83 p. Dissertação (Mestrado) — Universidade Federal de São Carlos - Programa de Pós-Graduação em Ciência da Computação - PPGCC, São Carlos, 2017. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/8961?show=full>. Citado nas páginas 41 e 42.

BONTCHEVA K., C. H. **Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation. In: Handbook of Semantic Web Technologies. Springer, Berlin, Heidelberg**. 2011. Disponível em: <https://doi.org/10.1007/978-3-540-92913-0_3>. Citado na página 38.

CASTRO, S. de. **Ontologias**. [S.l.]: Zahar, 2008. v. 1. Citado na página 27.

CAVALCANTI, A. P. **Uma medida de similaridade textual para identificação de plágio em fóruns educacionais**. 88 p. Dissertação (Mestrado) — Universidade Federal Rural de Pernambuco - Programa de Pós-Graduação em Informática Aplicada, Pernambuco, 2018. Disponível em: <http://www.tede2.ufrpe.br:8080/tede2/handle/tede2/7868>. Citado na página 38.

COELHO, R. da S. V. **Uma Abordagem para Anotação Semântica de Mídia Enativa para o Ensino de Conceitos de Matemática**. 109 p. Dissertação (Mestrado) — Centro Universitário de Campo Limpo Paulista - Programa de Mestrado Profissional em Ciência da Computação, Campo Limpo Paulista, 2022. Disponível em: <https://www.cc.faccamp.br/Dissertacoes/RaquelSilvaVieiraCoelho.pdf>. Citado na página 42.

CORRÊA, G. C.; MARCACINI, R. M.; REZENDE, S. O. **Uso da mineração de textos na análise exploratória de artigos científicos - Relatório Técnico, ICMC/USP.** [S.l.], 2012. Disponível em: <http://repositorio.icmc.usp.br/handle/RIICMC/6631>. Citado na página 39.

DAMASCENO, F. R.; RIBEIRO, A.; REATEGUI, E. Aplicação educacional de uma ferramenta de mineração de textos integrada a uma ontologia de domínio na Área da saúde. **RENOTE**, v. 9, n. 1, jul. 2011. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/21912>. Citado nas páginas 41 e 42.

DING, Y.; ENGELS, R. Ir and ai: Using co-occurrence theory to generate lightweight ontologies. **DEXA Workshop**, v. 1, p. 961–965, 2001. Disponível em: <https://ieeexplore.ieee.org/document/953179>. Citado na página 28.

FONTES, C. A. **EXPLORANDO INFERÊNCIA EM UM SISTEMA DE ANOTAÇÃO SEMÂNTICA**. 126 p. Dissertação (Mestrado) — Instituto Militar de Engenharia, Rio de Janeiro, 2011. Disponível em: <http://www.comp.ime.eb.br/pos/modules/files/dissertacoes/2011/2011-Celso.pdf>. Citado na página 65.

GODOIS, L. M. **Um algoritmo de agrupamento de dados utilizando interação entre agentes**. 57 p. Dissertação (Mestrado) — Universidade Federal do Rio Grande - Programa de Pós-graduação em Modelagem Computacional, Rio Grande, 2018. Disponível em: <https://repositorio.furg.br/handle/1/9170>. Citado nas páginas 39, 40, 41 e 42.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199–220, 1993. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1042814383710083>. Citado na página 27.

GUARINO, N. Formal ontology and information system. In: FOIS, 1. Trento, 1998. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=Wf5p3_fUxacC&oi=fnd&pg=PR5&ots=noWJXToEKN&sig=JUQappsFa9nRprSrjkDp53rY1Xg&redir_esc=y#v=onepage&q&f=false>. Citado na página 28.

GÓMEZ-PÉREZ, A. **Ontology Evaluation. Handbook on Ontologies.** Springer, 2004. v. 1. 251-274 p. Disponível em: <https://link.springer.com/content/pdf/10.1007/978-3-540-24750-0.pdf>. Citado na página 28.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, v. 4, p. 18–36, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>. Citado na página 39.

MARCONI, M. A.; LAKATOS, E. M. **Técnicas de pesquisa**. [S.l.]: Atlas, 1999. v. 1. Citado na página 21.

MENDONçA, F. M. **ONTOFORINFOSCIENCE: METODOLOGIA PARA CONSTRUÇÃO DE ONTOLOGIAS PELOS CIENTISTAS DA INFORMAÇÃO. Uma aplicação prática no desenvolvimento da ontologia sobre componentes do sangue humano (HEMONTO)**. Tese (Doutorado) — Universidade Federal de Minas Gerais - Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação, 2015. Disponível em: <https://repositorio.ufmg.br/handle/1843/BUBD-A35H3K>. Citado nas páginas 29, 32, 33, 34, 35, 36, 43, 54, 55, 62 e 63.

PEREIRA, J. W. **Anotação semântica baseada em ontologia: um estudo do português brasileiro em documentos históricos do final do século XIX**. 99 p. Dissertação (Mestrado) — Universidade Federal de São Carlos - Programa de Pós-graduação em Ciências da Computação, São Carlos, 2014. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/561>. Citado nas páginas 41, 42 e 65.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, v. 7, p. 7–21, 2011. ISSN 19835604. Disponível em: <http://www.fsma.edu.br/si/7edicao.html>. Citado nas páginas 38 e 39.

ROSINA, D. **Entre narrativas, fragmentos e estilhas: construções de atletas brasileiros sobre os jogos olímpicos do México de 1968**. 214 p. Dissertação (Mestrado) — Universidade de São Paulo - Escola de Educação Física e Esporte, São Paulo, 2018. Disponível em: <https://teses.usp.br/teses/disponiveis/39/39136/tde-20022019-100804/pt-br.php>. Citado na página 23.

RUBIO, K. **Memórias e narrativas biográficas de atletas olímpicos brasileiros. In: Rubio K, organizador. Preservação da memória: a responsabilidade social dos Jogos Olímpicos.** Képos, 2014. v. 1. Disponível em: <https://www.researchgate.net/publication/282862452_Preservacao_da_Memoria_a_responsabilidade_social_dos_Jogos_Olimpicos>. Citado na página 22.

____. **Atletas Olímpicos Brasileiros**. Sesi, 2015. v. 1. Disponível em: <https://www.researchgate.net/publication/281178503_Atletas_Olimpicos_Brasileiros>. Citado na página 22.

RUBIO, K.; RABELO, I. S.; SINOARA, R. A.; BARBOSA, R. S.; REZENDE, S. O. Identificação de personalidade de atletas olimpicos: uma análise exploratória de narrativa com mineração de textos. **REVISTA BRASILEIRA DE PSICOLOGIA DO ESPORTE**, v. 9, p. 1–19, 2019. Disponível em: <https://portalrevistas.ucb.br/index.php/RBPE/article/view/10568>. Citado nas páginas 24 e 42.

SANTOS, L. C. de M. **RECUPERAÇÃO DA INFORMAÇÃO EM ACERVOS DIGITAIS DE JORNAIS: proposta para uso de ontologia no domínio do futebol**. 201 p. Dissertação (Mestrado) — Universidade Federal de Santa Catarina - Programa de Pós-Graduação em Ciência da Informação, Santa Catarina, 2016. Disponível em: <https://repositorio.ufsc.br/handle/123456789/169099>. Citado nas páginas 41, 42, 55 e 56.

SCHIESSL, M.; BRäSCHER, M. Do texto às ontologias: uma perspectiva para a ciência da informação. **Ciência da Informação**, v. 40, p. 301–311, 2012. Disponível em: <https://revista.ibict.br/ciinf/article/view/1318>. Citado na página 30.

SILVA, W. D. da; PARREIRAS, F. S.; MAIA, L. C. G.; BRANDÃO, W. C. Anotação semântica automática do currículo lattes utilizando linked open data. **Perspectivas em Ciência da Informação**, v. 23, p. 53–72, 2018. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22591>. Citado na página 67.

TAN, A. H. Text mining: The state of the art and the challenges," proceedings of the pakdd workshop on knowledge discovery from advanced databases, beijing. In: . [s.n.], 1999. p. 65–70. Disponível em: <https://www.researchgate.net/profile/Ah-Hwee-Tan/publication/2471634_Text_Mining_The_state_of_the_art_and_the_challenges/links/54b924610cf269d8cbf73381/Text-Mining-The-state-of-the-art-and-the-challenges.pdf>. Citado na página 23.

ZIMMERMANN, M. A. **O professor inesquecível nas narrativas de atletas olímpicos brasileiros**. 116 p. Dissertação (Mestrado) — Universidade de São Paulo - Escola de Educação Física e Esporte, São Paulo, 2019. Disponível em: <https://www.teses.usp.br/teses/disponiveis/39/39136/tde-14062019-093228/pt-br.php>. Citado na página 23.

APÊNDICE

A

APÊNDICE ONTOLOGIA

A.1 Dicionário de conceitos

	Conceito	Sinônimos	Definição
1	Pessoa		Ser humano; quem pertence à espécie humana; criatura.
2	Atleta	Desportista	Pessoa treinada para competir, profissionalmente ou como amador, em exercícios, esportes ou jogos que requerem força, agilidade e resistência; esportista.
3	Família		Grupo de pessoas que partilha ou que já partilhou a mesma casa, normalmente estas pessoas possuem relações entre si de parentesco, de ancestralidade ou de afetividade.
4	Professor		Aquele que leciona em algum estabelecimento de ensino; docente, mestre, prô
5	Treinador	técnico	Que ou aquele que treina ou dirige um time esportivo; técnico
6	Transição	aposentadoria	No sentido do grupo de estudos olímpicos, transição trata do momento em que o atleta deixa de exercer atividade esportiva competitiva.
7	Profissão		Ocupação ou emprego do qual se obtém o sustento para si e seus dependentes
8	Saúde		Estado do organismo com funções fisiológicas regulares e com características estruturais normais e estáveis, levando-se em consideração a forma de vida e a fase do ciclo vital de cada ser ou indivíduo. Bem-estar físico, psíquico e social.
9	Mental		estado de equilíbrio mental de um indivíduo, adaptado ao seu meio social e bem tolerante às condições e desafios da existência social e individual.
10	Lesão		Ato ou efeito de lesar. Ferimento ou traumatismo.
11	Dor		Sensação desagradável ou penosa, de intensidade variável, causada por um estado anômalo do organismo ou parte dele e mediada pela estimulação de fibras nervosas que levam os impulsos dolorosos para o cérebro; sofrimento físico.

12	Peso		Medida da força gravitacional com que os corpos são atraídos para o ponto central da Terra.
13	Jogos Olímpicos		Evento poliesportivo realizado a cada quatro anos.
14	Sede		Centro ou ponto escolhido para nele se estabelecer alguma coisa
15	Modalidade		Aspecto, forma ou característica particular de algo; tipo. (Esporte)
16	Edição		Realização periódica de um evento (artístico, cultural, esportivo etc.)
17	Clube		Local com instalações para a prática de diversas modalidades de esportes e recreação
18	E-Sports		E-Sport é o termo utilizado para classificar competições de jogos virtuais, especialmente aquelas realizadas por profissionais, que podem ser assistidas pelo público pela televisão ou por plataformas de streaming.
19	Doping		Uso ilegal, por um atleta, de substâncias químicas que lhe aumentam o desempenho
20	Militar		Que se baseia e se apoia nas Forças Armadas
21	Formação		A educação acadêmica de um indivíduo, incluindo-se os cursos concluídos e os títulos obtidos
22	Curso		Programa educacional que tem o propósito de ensinar
23	Instituição		Estabelecimento de ensino
24	Nível		Nível de Ensino
25	Raça		Divisão dos vários grupos humanos, diferenciados uns dos outros por caracteres físicos hereditários, tais como a cor da pele, o formato do crânio, as feições, o tipo de cabelo etc., embora haja variações de indivíduo para indivíduo dentro do mesmo grupo. [A noção de raça é bastante discutível, pois deve-se considerar com mais relevância a proximidade cultural do que o aspecto racial.
26	Gênero		Categoria que indica, por meio de desinências, uma divisão dos nomes baseada em critérios tais como sexo e associações psicológicas
27	LGBTQIA++		lésbicas, gays, bissexuais, travestis, transsexuais, queer e outros grupos de gênero e sexualidade.
28	Mulheres		Ser humano do sexo feminino
29	Lugar		País, região, estado, cidade, povoado etc. não determinado
30	Cidade		Grande aglomeração de pessoas em uma área geográfica circunscrita, com

			inúmeras edificações, que desenvolve atividades sociais, econômicas, industriais, comerciais, culturais, administrativas etc.; urbe.
31	Estado		Cada um dos territórios de certos países
32	País		Território com delimitações geográficas definidas, habitado por uma coletividade, com história e cultura próprias

A.2 Tabela de valores

	Conceito	Sinônimos	Definição	Valores
1	Pessoa		Ser humano; quem pertence à espécie humana; criatura.	Atleta, Família, Professor, Treinador
2	Atleta	Desportista	Pessoa treinada para competir, profissionalmente ou como amador, em exercícios, esportes ou jogos que requerem força, agilidade e resistência; esportista.	
3	Família		Grupo de pessoas que partilha ou que já partilhou a mesma casa, normalmente estas pessoas possuem relações entre si de parentesco, de ancestralidade ou de afetividade.	Mãe, Pai, Irmão, Irmã, Avô, Avó, Esposa, Marido, Tia, Tio, Prima, Primo, Sobrinho, Sobrinha, Neto, Neta.
4	Professor		Aquele que leciona em algum estabelecimento de ensino; docente, mestre, prô	
5	Treinador	técnico	Que ou aquele que treina ou dirige um time esportivo; técnico	
6	Transição	Aposentadoria	No sentido do grupo de estudos olímpicos, transição trata do momento em que o atleta deixa de exercer atividade esportiva competitiva.	

7	Profissão		Ocupação ou emprego do qual se obtém o sustento para si e seus dependentes	Administrador, Administradora Advogada, Advogado, Aposentado, Arquiteta, Arquiteto, Babá, Consultor Consultora, Contador Contadora, Cozinheira Cozinheiro, Dentista Desempregada, Desempregado, DonaDeCasa Economista, EducadoraFísica EducadorFísico, Empresária Empresário, Enfermeira Enfermeiro, Engenheira Engenheiro, Estagiária Estagiário, Estudante Fisioterapeuta, Gestor Gestora, Guarda Jornalista, Motorista Médica, Médico Pedreira, Pedreiro Policial, Política Político, PreparadoraFísico PreparadorFísico, ProfessoraDeEducaçãoFísica ProfessorDeEducaçãoFísica Psicóloga, Psicólogo, Treinador, Treinadora
8	Saúde		Estado do organismo com funções fisiológicas regulares e com características estruturais normais e estáveis, levando-se em consideração a forma de vida e a fase do ciclo vital de cada ser ou indivíduo. Bem-estar físico, psíquico e social.	Mental, Lesões, Dor, Peso
9	Mental		estado de equilíbrio mental de um indivíduo, adaptado ao seu meio social e bem tolerante às condições e desafios da existência social e individual.	sofre Sofremos sofrimentos suicídio traumatizada travada treme Tremem Tremendo Depressão Ansiedade
10	Lesão		Ato ou efeito de lesar. Ferimento ou traumatismo.	Coluna, Cotovelo, Dor, Joelho, ombro, Tendão, Tibiais, Tornozelo, trincada, trincadinha, trincou, tíbias Machucou, machucar

11	Dor		Sensação desagradável ou penosa, de intensidade variável, causada por um estado anômalo do organismo ou parte dele e mediada pela estimulação de fibras nervosas que levam os impulsos dolorosos para o cérebro; sofrimento físico.	Dolorido, Doloroso, dorzinha, Incômodo
12	Peso		Medida da força gravitacional com que os corpos são atraídos para o ponto central da Terra.	Anorexia, Bulimia, Dieta Gorda, gordo, Obesa, Obeso, Regime, Vomitava, Vômito
13	Jogos Olímpicos		Evento poliesportivo realizado a cada quatro anos.	Edição, Sede e Modalidade
14	Sede		Centro ou ponto escolhido para nele se estabelecer alguma coisa	Antuérpia, Atenas, Atlanta, Barcelona, Berlim, Brisbane, CidadeDoMéxico, Estocolmo, Helsinque, Londres, LosAngeles, Melbourne, Montreal, Moscou, Munique, Paris, Pequim, RioDeJaneiro, Roma, SaintLouis, Seul, Sydney, Tóquio
15	Modalidade		Aspecto, forma ou característica particular de algo; tipo. (Esporte)	Atletismo, Badminton, Basquetebol, Boxe, Canoagem, Ciclismo, Esgrima, Futebol, GinásticaArtística, GinásticaRítmica, GinásticaTrampolim, Golfe, Halterofilismo, Handebol, Hipismo, HóqueiSobreGrama, Iatismo, Judô, Lutas, NadoArtístico, NadoSincronizado, Natação, PentatloModerno, PoloAquático, Remo, Rugby, SaltosOrnamentais, Taekwondo, Tiro, TiroComArco, Triatlo, Tênis, TênisDeMesa, Vela, VoleibolDePraia, VoleibolDeQuadra,
16	Edição		Realização periódica de um evento (artístico, cultural, esportivo etc.)	1896, 1900, 1904, 1908, 1912, 1920, 1924, 1928, 1932, 1936, 1948, 1952, 1956, 1960, 1964, 1968, 1972, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020, 2024, 2028, 2032
17	Clube		Local com instalações para a prática de diversas modalidades de	ABC, AMERICANO, AMÉRICA-MG, AMÉRICA-PE, AMÉRICA-RJ, AMÉRICA-RN, AMÉRICADERIOPRETO, ANAPOLINA, ANDRADINA, ARAÇATUBA,

			esportes e recreação	ATHLETICO-PR, ATLÉTICO-GO, ATLÉTICO-MG, ATLÉTICOSOROCABA, AVAÍ, AVENIDA, BAHIA, BAHIADEFEIRA, Bangu, BARRAS, BARRETOS, BATATAIS, BAURUA.C., BOTAFOGO, BOTAFOGO-BA, BOTAFOGO-PB, BOTAFOGODERIBEIRÃOPRETP, BRASILIENSE, BRASÍLIA, BRUSQUE, CALDENSE, CAMBARAENSE, CAMPINENSE, CAPIVARIANO, CAXIAS, CEARÁ, CENE, CEUB, CHAPECOENSE, CIANORTE, ClubeGuarulhos, ClubeMaranhão, ClubeOsasco, COMERCIAL, COMERCIALDERIBEIRÃOPRETO, CONFIANÇA, Corinthians, CORITIBA, CORUMBAENSE, CRB, CRICIÚMA, CRUZEIRO, Cruzeiro, CSA, CUIABÁ, CírculoMilitar, DESPORTIVA, DOMBOSCO, DOURADOS, EsporteClubePinheiros, Espéria, FAST, FERROVIÁRIO, FIGUEIRENSE, FLAMENGO, FLAMENGO-PI, Fluminense, FLUMINENSEDEFEIRADESANTANA, FORTALEZA, Franca, FRANCANA, GAMA, GERMÂNIA, GOIÁS, GOIÂNIA, Grêmio, GRÊMIO, GRÊMIOBARUERI, GRÊMIOCATANDUVENSEDEFUTEBOL, GRÊMIONOVORIZONTINO, GRÊMIOOSASCO, GRÊMIOOSASCOAUDAX, GRÊMIOPRUDENTE, GUARANI, GUARATINGUETÁ, GUARÁ, Hebraica, INTERNACIONAL, INTERNACIONALDELIMEIRA, ITABAIANA, ITUANO, ITUMBIARA, JABAQUARA, JOINVILLE, JUVENTUDE, LEÔNICO, LINENSE, LONDRINAE.C., LUVERDENSE, MADUREIRA, MARINGÁF.C., MARÍLIA, MATONENSE, Metodista, MinasTênisClube, MIRASSOL, MISTO, MIXTO, MOGI-MIRIM, MOGIANA, MONTEALEGRE, MONTEAZUL, MonteLíbano, MOTOCLUBE, NACIONAL, NACIONAL-AM, NOROESTE, NOVORIZONTINO, NÁUTICO, OESTE, OLARIA, OPERÁRIODECAMPOGRANDE, Paineiras, PALMEIRAS, PARANÁ, PATOBRANCO, PAULISTADEJUNDIAÍ, PAULISTANO, PAYSANDU-PA, PENAPOLENSE, PONTEPRETA, PORTUGUESA, PORTUGUESA-RJ, PORTUGUESASANTISTA, PRESIDENTEPRUDENTE, RADIUM, REDBULLBRAGANTINO,

				REDBULLBRASIL, REMO-PA, RETRÔ, RibeirãoPreto, RIOBRANCO-AC, RIOBRANCO-ES, RIOBRANCODEAMERICANA, RIOCLARO, RIONEGRO, RIOPRETO, RioYachtClube, RIVER, SAAD, SALGUEIROFutebol, SAMPAIOCORRÊA, SANTACRUZ, SANTOANDRÉ, SANTOS, SantosFutebolClube, SERGIPE, SERTÃOZINHO, Sesi, SOBRADINHO, SOGIPA, SPORT, Sport, SUZANO, SÃOBENTODESOROCABA, SÃOCAETANO, SÃOCARLENSE, SÃOCRISTÓVÃOF.R., SÃOJOSÉE.C., SãoPauloFutebolClube, SÍRIO, TAQUARITINGA, TAUBATÉ, Tietê, TijucaTênisClube, TIRADENTES, TREZE, TUPÃ, UBERLÂNDIA, UMUARAMA, UNIÃOBARBARENSE, UNIÃODEMOGI, UNIÃOSÃOJOÃODEARARAS, VASCO, VascoDaGama, VELOCLUB, VILANOVA-GO, VILLANOVA-MG, Vitória, VITÓRIA, VOCEM, VOTORANTIM, VOTUPORANGUENSE, XVDEJAÚ, XVDEPIRACICABA, YPIRANGA, ÁGUASANTA
18	Esports			Games, Jogos, eletrônicos
19	Doping		Uso ilegal, por um atleta, de substâncias químicas que lhe aumentam o desempenho	Banido, Dopado, Suspenso, Suspensão
20	Militar		Que se baseia e se apoia nas Forças Armadas	Exército, marinha, aeronáutica
21	Formação		A educação acadêmica de um indivíduo, incluindo-se os cursos concluídos e os títulos obtidos	Curso, Instituição e Nível
22	Curso		Programa educacional que tem o propósito de ensinar	Administração, Direito, Economia, Educação_Física, Enfermagem, Fisioterapia, Gestão, Marketing, Medicina, Odontologia
23	Instituição		Estabelecimento de ensino	Mackenzie, PUC, UFMG, UFRJ, Unesp, Unicamp, UNIP, Unisanta, UniSantana, UNOPAR, USP
24	Nível		Nível de Ensino	
25	Raça		Divisão dos vários grupos humanos, diferenciados uns dos outros por caracteres físicos	Negro, Negritude, racismo, racial

			hereditários, tais como a cor da pele, o formato do crânio, as feições, o tipo de cabelo etc., embora haja variações de indivíduo para indivíduo dentro do mesmo grupo. [A noção de raça é bastante discutível, pois deve-se considerar com mais relevância a proximidade cultural do que o aspecto racial.	
26	Gênero		Categoria que indica, por meio de desinências, uma divisão dos nomes baseada em critérios tais como sexo e associações psicológicas	
27	LGBTQI++		lésbicas, gays, bissexuais, travestis, transsexuais, queer e outros grupos de gênero e sexualidade.	lésbicas, gays, bissexuais, travestis, transsexuais, queer, trans
28	Mulher		Ser humano do sexo feminino	
29	Lugar		País, região, estado, cidade, povoado etc. não determinado	Cidade, Estado, País
30	Cidade		Grande aglomeração de pessoas em uma área geográfica circunscrita, com inúmeras edificações, que desenvolve atividades sociais, econômicas, industriais, comerciais, culturais, administrativas etc.; urbe.	Abidjan, AbuDhabi, Abuja, Acra, AdisAbeba, Americana, Amã, Anapolina, Ancara, Andorra-a-Velha, Andradina, Antananarivo, Antuérpia, Apia, Aracajú, Araras, Araçatuba, Argel, Asgabate, Asmara, Assis, Assunção, Astana, Atenas, Atlanta, Bagdá, Baku, Bamaco, BandarSeriBegauã, Banguecoque, Bangui, Banjul, Barcelona, Barras, Barretos, Barueri, Basseterre, Batatais, Bauru, Beirute, Belgrado, Belmopã, BeloHorizonte, Belém, Berlim, Berna, Bisqueque, Bissau, BoaVista, Bogotá,

				BragançaPaulista, Brasília, Bratislava, Brazavile, Bridgetown, Brisbane, Brusque, Bruxelas, Bucareste, Budapeste, BuenosAires, Bujumbura, Cabul, Cairo, Camaragibe, Cambará, Camberra, Campala, CampinaGrande, Campinas, CampoGrande, Capivari, Caracas, Cartum, Castries, Catanduva, Caxias, Chapecó, Cianorte, CidadeDaGuatemala, CidadeDoKwait, CidadeDoMéxico, CidadeDoPanamá, Conacri, Copenhaguen, Corumbá, Criciúma, Cuiabá, Curitiba, Daca, Dacar, Damasco, Diadema, Djibuti, Dodoma, Doha, Dourados, Dublin, Duchambé, Díli, Escópia, Estocolmo, FeiraDeSantana, Florianópolis, Fortaleza, Franca, Freetown, Funafuti, Gaborone, Georgetown, Goiânia, Guaratinguetá, Guarulhos, Hanói, Harare, Havana, Helsinque, Honiara, Iarém, Iaundé, Islamabad, Itabaiana, Itu, Itumbiara, Itápolis, Jacarta, Jamena, Jaú, Jerusalém, JerusalémOriental, Joinville, JoãoPessoa, Juba, Jundiaí, Katmandu, Kingston, Kingstown, KualaLumpur, LaPaz, Libreville, Lilôngue, Lima, Limeira, Lins, Lisboa, Liubliana, Lobamba, Lomé, Londres, Londrina, LosAngeles, Luanda, LucasDoRioVerde, Lusaca, Luxemburgo, Macapá, Maceió, Madrid, Malabo, Malé, Manama, Manaus, Manila, Manágua, Maputo, Maringá, Marília, Mascate, Maseru, Matão, Melbourne, Minsk, Mirassol, Mococa, Mogadíscio, MogiDasCruzes, MogiMirim, Monróvia, MonteAlegre, MonteAzulPaulista, Montevideu, Montreal, Moroni, Moscou, Moscovo, Munique, Mônaco, Nairóbi, Nassau, Natal, Naypyidaw, Ngerulmud, Niamei, Nicósia, Nouakchott, NovaDéli, NovaLima, NovoHorizonte, Nucualofa, Osasco, Oslo, Otawa, Paliquir, Palmas, Paramaribo, Paris, PatoBranco, Penápolis, Pequim, Pionguiangue,

				Piracicaba, PnomPene, Podgoritsa, PortoAlegre, PortoDeEspanha, PortoLuís, PortoMoresby, PortoNovo, PortoPríncipe, PortoVelho, PortoVila, PoçosDeCaldas, Praga, Praia, PresidentePrudente, Pretória, Pristina, Quieve, Quigali, Quinxassa, Quito, Quixinau, Rabat, Recife, Reykjavik, Riad, RibeirãoPreto, Riga, RioBranco, RioClaro, RioDeJaneiro, RioPreto, Roma, Roseau, SaintLouis, Salgueiro, Salvador, SantaBárbaraDoOeste, SantaCruzDoSul, Santiago, SantoAndré, SantoDomingo, Santos, Saná, Sarajevo, Sertãozinho, Seul, Singapura, Sorocaba, SriJaiavardenapura-Cota, Suva, Suzano, Sydney, SãoBernardoDoCampo, SãoCaetano, SãoCarlos, SãoJorge, SãoJosé, SãoJoséDosCampos, SãoJoão, SãoLuis, SãoLuís, SãoMarinho, SãoPaulo, SãoSalvador, SãoTomé, Sófia, Taipé, Talim, Taquaritinga, TarauaDoSul, Tasquente, Taubaté, Tebilíssi, Teerão, Tegucigalpa, Teresina, Timbu, Tirana, Trípoli, Tunes, Tupã, Tóquio, Uagadugu, Uberlândia, UlanBator, Umuarama, Vaduz, Valeta, Varsóvia, Vaticano, Viena, Vienciana, Vinduque, Vitória, Votorantim, Votuporanga, Vílnius, Wellington, Yerevã, Zagreb
31	Estado		Cada um dos territórios de certos países	Acre, Alagoas, Amapá, Amazonas, Bahia, Ceará, EspiritoSanto, EspíritoSanto, Goiás, Maranhão, MatoGrosso, MatoGrossoDoSul, MinasGerais, Paraná, Paraíba, Pará, Pernambuco, Piauí, Rio_de_Janeiro, RioDeJaneiro, RioGrandeDoNorte, RioGrandeDoSul, Rondônia, Roraima, SantaCatarina, Sergipe, SãoPaulo, Tocantins
32	País		Território com delimitações geográficas definidas, habitado por uma	Afeganistão, Albânia, Alemanha, Andorra, Angola, AntigaEBarbuda, Argentina, Argélia, Armênia, ArábiaSaudita, Austrália, Azerbaijão, Bahamas, Bahrein, Bangladesh,

			coletividade, com história e cultura próprias	Barbados, Belize, Benim, Bielorrússia, Bolívia, Botsuana, Brasil, Brunei, Bulgária, BurkinaFaso, Burúndi, Butão, Bélgica, BósniaEHerzegovina, CaboVerde, Camarões, Camboja, Canadá, Catar, Cazaquistão, Chade, Chile, China, Chipre, Colômbia, Comores, Congo, CoreiaDoNorte, CoreiaDoSul, CostaDoMarfim, CostaRica, Croácia, Cuba, Dinamarca, DJibuti, Dominica, Egito, ElSalvador, EmiradosÁrabesUnidos, Equador, Eritreia, Eslováquia, Eslovênia, Espanha, Essuatíni, EstadosUnidos, Estônia, Etiópia, EUA, Fiji, Filipinas, Finlândia, França, Gabão, Gana, Geórgia, Granada, Grécia, Guatemala, Guiana, Guiné, Guiné-Bissau, GuinéEquatorial, Gâmbia, Haiti, Holanda, Honduras, Hungria, IlhasMaurício, IlhasSalomão, Indonésia, Iraque, Irlanda, Irã, Islândia, Israel, Itália, Iêmen, Jamaica, Japão, Jordânia, Kiribati, Kosovo, Kwait, Laos, Lesoto, Letônia, Libéria, Liechtenstein, Lituânia, Luxemburgo, Líbano, Líbia, MacedôniaDoNorte, Madagascar, Malawi, Maldivas, Mali, Malta, Malásia, Marrocos, Mauritânia, Mianmar, Micronésia, Moldávia, Mongólia, Montenegro, Moçambique, México, Mônaco, Namíbia, Nauru, Nepal, Nicarágua, Nigéria, Noruega, NovaZelândia, Níger, Omã, Palau, Palestina, Panamá, PapuaNovaGuiné, Paquistão, Paraguai, Peru, Polônia, Portugal, Quirguistão, Quénia, ReinoUnido, RepúblicaCentro-Africana, RepúblicaDemocráticaDoCongo, RepúblicaDominicana, RepúblicaTcheca, Romênia, Ruanda, Rússia, Samoa, SantaLúcia, Seicheles, Senegal, SerraLeoa, Singapura, Somália, SriLanka, Sudão, SudãoDoSul, Suriname, Suécia, Suíça, SãoCristóvãoENeves, SãoMarinho, SãoToméEPríncipe, SãoVicenteEGranadinas, Sérvia,

				Síria, Tailândia, Taiwan, Tajiquistão, Tanzânia, Timor-Leste, Togo, Tonga, TrindadeETobago, Tunísia, Turcomenistão, Turquia, Tuvalu, Ucrânia, Uganda, UniãoSoviética, Uruguai, Uzbequistão, Vanuatu, Vaticano, Venezuela, Vietnã, Zimbábue, Zâmbia, ÁfricaDoSul, Áustria, Índia


ICMC
USP
SÃO CARLOS