Universidade Federal do Rio de Janeiro

MIGUEL GABRIEL PRAZERES DE CARVALHO

ENRIQUECIMENTO DE ONTOLOGIAS: UMA ABORDAGEM PARA EXTRAÇÃO DE INFORMAÇÕES IMPLÍCITAS

DISSERTAÇÃO DE MESTRADO

Instituto de Matemática

Núcleo de Computação Eletrônica

MIGUEL GABRIEL PRAZERES DE CARVALHO

Enriquecimento de Ontologias: Uma Abordagem para Extração de Informações Implícitas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientadores:

Maria Luiza Machado Campos, Ph.D.

Vanessa Braganholo Murta, D.Sc.

Maria Luiza de Almeida Campos, D.Sc.

Rio de Janeiro

2011

C331 Carvalho, Miguel Gabriel Prazeres de
Enriquecimento de ontologias: uma abordagem para extração de informações implícitas / Miguel Gabriel Prazeres de Carvalho –, 2011.
147 f.; il.
Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Núcleo de Computação Eletrônica, Programa de Pós-Graduação em Informática, 2011.
Orientadores: Maria Luiza Machado Campos, Vanessa Braganholo Murta e Maria Luiza de Almeida Campos
1. Ontologias. 2. Bioinformática. 3. Extração de Informação. 4. Informações Implícitas. 5. Reúso de Ontologias. 6. Alinhamento. 7. Consórcio OBO - Teses. I. Maria Luiza Machado Campos (Orient.) II. Vanessa Braganholo Murta (Orient.) III. Maria Luiza de Almeida Campos (Orient.) IV. Título.
CDD

MIGUEL GABRIEL PRAZERES DE CARVALHO

Enriquecimento de Ontologias: Uma Abordagem para Extração de Informações Implícitas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Aprovado em: Rio de Janeiro, 25 de março de 2011.

_____

Professora Maria Luiza Machado Campos, Ph.D., PPGI/UFRJ (Orientadora)

_____

Professora Vanessa Braganholo Murta, D.Sc., PPGI/UFRJ (Orientadora)

_____

Professora Maria Luiza de Almeida Campos, D.Sc., PPGCI/UFF (Orientadora)

_____

Professor João Carlos Pereira da Silva, D.Sc., PPGI/UFRJ

_____

Professora Jonice de Oliveira Sampaio, D.Sc., PPGI/UFRJ

_____

Professora Ana Maria de Carvalho Moura, Dr.Ing., LNCC/RJ

_____

Professora Hagar Espanha Gomes, Livre-docência, CNPq

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus, Nossa Senhora e Santa Rita de Cássia por terem me permitido chegar até aqui, concluindo mais uma etapa da minha vida, me dando força, esperança e coragem nos momentos mais difíceis e cuidando de mim todos os dias. Posteriormente, gostaria de agradecer aos presentes que ganhei de Deus: minha família, meus amigos e meus professores.

Gostaria de agradecer, primeiro, aos melhores e mais especiais presentes que Deus me deu: meu pai Antônio, minha mãe Lila e minha irmã Clara por todo amor, apoio e incentivo durante toda a minha vida.

Aos meus avôs: Adélia e Ivo, Ruben (em memória) e Vera, José Valouis (em memória) e Elza (em memória).

Aos meus tios e primos: Verônica, Luis, Pedro, Maria, Luizinho, Ana, Gabriel, Gustavo, Frederico, Felipe, Barbara, Ary, Binho.

Ao meu cunhado Marcelo Reis, o irmão que eu não tive e um amigo que carregarei para vida inteira.

À minha amada namorada Andressa pela amizade, amor, incentivo e entendimento em todos os momentos.

Agradeço ao meu amigo-irmão Rubinho por toda a amizade, companheirismo e também toda a ajuda na parte de entendimento do framework Jena. A minha amiga Vivian por toda a paciência e auxílio na parte de avaliação experimental e também no entendimento de diversos conceitos da área biomédica. Ao meu amigo Fabrício, pelas aulas de ferramentas de extração de análise linguísticas. Aos amigos biólogos Nicole e Diogo por me ajudarem na validação da avaliação experimental.

Aos meus amigos Linair, Serra, João Vitor, Inês, Eliêmia, Patricia, Wendel, Antônio (Montanha), Alex Sanches, Fábio, Maria Célia, Kelli, Bianca, Renan, Rafael Escalfoni, Bruno Caputo, Alan, Fernanda, Antônio e André companheiros constantes nessa jornada.

A todos os meus amigos que me ajudaram, diretamente ou indiretamente, nessa dissertação, em especial: Doutor Bruno, Roginaldo Reis, Verônica Reis, Ana Carolina, Hayla, Andréa Shirlei, Guilherme, Jorginho, Jorge Luis, Maria Izabel, Cristiano, Claudio

Micele, Bianca, Andre Abrantes, Annete Perorazio, Bruno Moura, Sérgio (Baiano), Célia, Ariane, Lilian, Márcio, Mônica, Eduardo, Florinda, Adilson, Silvia, Marcelo, Lindaci, Thiago Maluf, Bruno, Alexandre Sardinha, Douglas, Renato, Rafael Espírito, Carlos Eduardo, Carlos Henrique, Vitor, Thiago Berne, Thiago Nascimento, Carlos, Marcinha, Thiago, Anibal, Frank, Bruno Milet, Cacau, Cassius e outros que por ventura me esqueci nesse momento.

Aos Professores que foram grandes amigos me permitindo chegar até aqui:

À professora Vanessa: Exemplo de dedicação, entusiasmo e profissionalismo. Agradeço a ela por toda a compreensão, amizade, paciência, apoio e pelo imenso conhecimento transmitido. Uma pessoa por quem tenho grande admiração e respeito, e uma grande amiga que sei que levarei para a vida inteira. Agradeço também pela brilhante orientação.

À professora Maria Luiza (UFRJ): Uma verdadeira mãe para mim, que indiscutivelmente é uma das melhores pessoas que já conheci em minha vida, sempre me ajudando em todos os momentos que eu precisei me dando conselhos e incentivos, uma pessoa que admiro muito e tenho grande respeito. Agradeço também pela brilhante orientação.

À professora Maria Luiza (UFF): Por todo o incentivo, paciência, amizade e apoio que vão desde a iniciação científica a qual motivou a escolha por esse tema. Agradeço muito a ela por toda a ajuda no entendimento dos conceitos da Ciência da Informação que foram a base desse trabalho. Agradeço também pela brilhante orientação.

À professora Jonice: Pelo carisma e pelo sorriso amigo, uma pessoa impar que nunca se nega a ajudar ninguém. Uma amiga de coração. Agradeço a ela também por ter aceitado tão gentilmente fazer parte dessa banca.

Ao Professor Marcos Borges: Por toda a amizade e apoio nessa jornada, por todas as oportunidades e portas abertas por ele para mim. Uma pessoa que tenho muito orgulho em dizer que sou seu amigo e que jamais esquecerei.

Ao Professor João Carlos: Meu quarto orientador, devo a ele toda a ajuda na parte de inteligência artificial e algoritmos. Muito mais que um professor, um verdadeiro amigo, agradeço muito de coração por todas as horas extras despendidas em prol de me ajudar. Agradeço também por tão gentilmente aceitar o convite para participar dessa banca.

Ao Professor Ageu: Por toda a amizade e apoio durante essa minha jornada acadêmica

que começou com seu apoio e incentivo quando fui seu monitor. Um grande amigo que indiscutivelmente levarei para sempre.

À Professora Linair: Minha quinta orientadora e grande amiga se cheguei até aqui devo muito a ela por todos os conselhos, ajuda e dicas para minha dissertação e para minha vida.

À Professora Ana Maria: agradeço a ela pelos ensinamentos passados durante o mestrado e utilizados nessa dissertação, agradeço também por tão gentilmente aceitar o convite para participar dessa banca.

À Professora Hagar Espanha: Uma das precursoras desse trabalho juntamente com a Profa. Maria Luiza de Almeida Campos. Agradeço a ela por todo o conhecimento transmitido e também por tão gentilmente aceitar participar dessa banca.

Ao Professor Mitre: Pela amizade, pela ajuda e pelas conversas. Uma pessoa impar com qual sei que poderei contar sempre que precisar.

Também não posso deixar de citar os professores: Collier, Miguel Jonahtan, Austeclynio, Adriano, Adriana, Leonardo Murta, Maria Cláudia, Claudinha, Amauri, Pedro Manuel, Cabral, Eber, Marcos Elia, Jayme, Fabio Protti, Juliana, Aguiar e José Orlando por toda amizade e aprendizado durante a graduação e o mestrado.

Aos amigos da secretaria: Querida tia Daise, Tia Edileuza, Zezé, Patrick, Sandro, Simone.

Aos amigos da secretaria de pós-graduação: Anibal e Daniella.

Aos amigos da biblioteca: Selma Mendes, Luiza Linhares, Maria Aparecida, Maria das Graças e Silma Ribeiro.

Aos amigos da portaria: Carlos, João, Francisco, Leivaldo, Marcondes, Netwon, Raimundo, Robson e Sebastião.

Dedico essa dissertação a todos vocês. Meus amigos, muito obrigado.

Miguel Gabriel

RESUMO

Carvalho, Miguel Gabriel Prazeres. **Enriquecimento de Ontologias: Uma Abordagem para Extração de Informações Implícitas.** Dissertação (Mestrado em Informática) - Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.

As pesquisas das áreas biológica e biomédica, impulsionadas principalmente pela pesquisa Genômica, fizeram com que fosse produzido um grande volume de informações. Nesse contexto as ontologias vêm sendo utilizadas para representar esse domínio servindo como um importante apoio em diversas pesquisas, atuando, por exemplo, como auxílio para anotação de recursos genômicos. Atualmente as ontologias mais utilizadas são as do consórcio OBO (*Open Biomedical Ontologies*), principalmente a *Gene Ontology*. Entretanto grande parte dessas ontologias possui diversos problemas em sua construção e estruturação como, por exemplo, problemas relacionados à hierarquia dos termos, às relações entre eles e às definição associadas, dentre outros. A maioria desses problemas interfere diretamente nas pesquisas que utilizam essas ontologias. Contudo, também foi verificado que a nomenclatura dos termos e o campo definição contêm informações implícitas que, se explicitadas nessas ontologias, poderiam ajudar a resolver alguns dos problemas encontrados através do enriquecimento do conhecimento expresso. Para isso, no escopo deste trabalho, foi criada uma abordagem para extração de informação cujo foco principal é a extração de informações implícitas e a avaliação da consistência nas ontologias a fim de melhorar a qualidade dos processos envolvidos na sua utilização, principalmente o reúso e a anotação. Essa abordagem difere das demais por possuir uma macroetapa de conhecimento do corpus, anterior ao processo de extração propriamente dito e também por ser bastante flexível para ser adaptada a diversos contextos. No contexto deste trabalho a abordagem foi aplicada com sucesso a duas ontologias biomédicas do consórcio OBO para a extração de relações hierárquicas do tipo é_um (*is_a*). Sua aplicação nessas ontologias revelou algumas inconsistências que deveriam ser corrigidas e também a omissão de termos e relações que deveriam existir. Além disso, o uso dessas informações implícitas junto às ontologias pesquisadas fez com que fosse evidenciada uma melhoria no processo de alinhamento dessas ontologias.

Palavras-Chaves: Ontologias, Extração de Informação, Informações Implícitas, Bioinformática, Reúso, Campo Definição, Hierarquia dos Termos.

ABSTRACT

Carvalho, Miguel Gabriel Prazeres. **Enriquecimento de Ontologias: Uma Abordagem para Extração de Informações Implícitas.** Dissertação (Mestrado em Informática) - Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.

Research in the biological and biomedical fields, mainly driven by Genomic Research, produces a large amount of information. In this context, ontologies have been used to represent this domain serving as an important support in several researches, acting, for example, as an aid to annotation of genomic features. Currently, most ontologies used are those of the OBO (Open Biomedical Ontologies) consortium, especially the Gene Ontology. However, these ontologies have many problems in their construction and structure, such as, for example, problems related to the hierarchy of terms, to relationships between them and to the associated definition. Most of these problems directly interfere in researches that are using these ontologies. However, it was also observed that the nomenclature of the terms and the definition field contain implicit information which, if explicitly expressed, could help solve some of the problems, by enriching the knowledge expressed. For this, in the scope of this work, we created an approach for information extraction, whose primary focus is on the extraction of implicit information and on the evaluation of the consistency of ontologies to improve the quality of all processes involved in their use, particularly reuse and annotation. This approach differs from others by having a macro step for knowing the corpus, which is executed previously to information extraction process, that is flexible enough to adapt to different domains. In the context of this work this approach has been successfully applied to two biomedical ontologies of the OBO Consortium for extracting relationships of the type "is_a". It was able to reveal some inconsistencies that should be corrected and also the omission of terms and relations that should exist in the ontology. Moreover, as showed by our experimental results, the use of this implicit information has improved the alignment of these ontologies.

Key Words: Ontology, Information Extraction, Information Implied, Bioinformatics, Reuse, Field Definition, Hierarchy of Terms.

LISTA DE ABREVIATURAS

ANNIE - *A Nearly New Information Extraction System*

BP - *Biological Process*

GATE - *General Architecture for Text Engineering*

GO - *Gene Ontology*

INOH - *Integrating Network Objects with Hierarchies*

NLTK - *Natural Language Toolkit*

OBO - *Open Biomedical Ontologies/ Open Biological and Biomedical Ontologies*

OWL - *Ontology Web Language*

RO - *Relation Ontology*

UMLS - *Unified Medical Language System*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1: INTRODUÇÃO

Este capítulo contextualiza a dissertação apresentando a caracterização do problema e as motivações que levaram a este trabalho. Posteriormente apresenta os objetivos do trabalho desenvolvido nessa dissertação e o enfoque de solução. Ao final descreve a estruturação dos capítulos, resumindo o que será tratado por cada um.

No decorrer dos últimos anos as pesquisas ligadas às áreas biológicas e biomédicas geraram um grande volume de informações. Nessa dinâmica, o campo da Bioinformática, o qual tem por objetivo aplicar conhecimento da computação às áreas ligadas à Biologia e à Biomedicina, foi um dos que mais evoluíram. Esse campo evoluiu principalmente através das pesquisas na área Genômica, que fizeram crescer a necessidade por sistemas capazes de apoiar as inúmeras pesquisas dessa área.

Nesse processo de evolução, área de Bioinformática vem se utilizando cada vez mais de ontologias para auxiliar na representação dos domínios biológicos e biomédicos. Ontologias são estruturas que permitem a representação formal de um domínio do conhecimento, permitindo o entendimento dos conceitos do domínio por seres humanos e a inferência de conhecimento por máquinas. Atualmente, as ontologias vêm sendo aplicadas a diversas áreas de conhecimento como: Engenharia de Software (NARDI; FALBO, 2006), Banco de Dados (DEGGAU; FILETO, 2009), Web Semântica (PICKLER, 2007), Gestão do Conhecimento (GAVA; MENEZES, 2003) e muitas outras.

Por permitirem a representação formal de um domínio, na área biomédica, as ontologias se tornaram um importante instrumento de apoio para anotação de recursos genômicos, sendo um vocabulário controlado padronizado para a anotação de sequências, podendo atuar em todas as fases do processo de anotação (SILVA, 2010). O processo de anotação genômica compreende as atividades: (i) registrar recebimento da sequência, (ii) realizar a pré-análise, (iii) fazer a análise das sequências e a anotação e (iv) consolidar a anotação (BELLOZE, 2007). Ao término desse processo as anotações passam a compor grandes bases de anotações genômicas, permitindo o compartilhamento de informações, beneficiando as pesquisas em diversas áreas como agricultura, pecuária e saúde (BELLOZE, 2007; SILVA, 2010).

No processo de anotação, a ontologia mais utilizada atualmente é a *Gene Ontology* (GO) (2010a). A *Gene Ontology* é uma ontologia do domínio biomédico, constituída por termos, definições e relações sendo formada por três ramos: Função Molecular (*Molecular*

Function), Processo Biológico (*Biological Process*) e Componente Celular (*Cellular Component*) (LEWIS, 2004).

Além da GO, existem diversas outras ontologias, com diferentes temáticas, utilizadas para representação do domínio biomédico. Grande parte dessas ontologias faz parte do consórcio *Open Biomedical Ontologies* (OBO) (2010a). O consórcio OBO é um esforço colaborativo entre diversos grupos, a exemplo de: *Biomedical Informatics Research Network* (BIRN) (2010), *Berkeley Bioinformatics Open Source Project* (BBOP) (2010), *Immunology Database and Analysis Portal* (ImmPort) (2010), *National Center for Biomedical Ontology* (NCBO) (2010) e *Ontology Research Group* (ORG) (2010), entre outros. Esse consórcio tem como objetivo a definição de normas e padrões para criação, manutenção e reformulação de ontologias ortogonais e interoperáveis para uso compartilhado em diferentes domínios biomédicos (SMITH *et al.*, 2007; OBO, 2010a).

Embora importantes em alguns contextos, a fragmentação do conhecimento biomédico em diversas ontologias (algumas mais especializadas em determinados temáticas e outras de usos mais gerais) pode também trazer malefícios para um entendimento global do domínio. Nesse cenário, o reúso de ontologias, principalmente através do alinhamento, pode melhorar significativamente o entendimento de um domínio, uma vez que permite a utilização de diferentes ontologias em conjunto, sem que seja necessário ao pesquisador analisar cada uma individualmente (SILVA, 2010). Através do reúso de ontologias é possível melhorar, entre outras coisas, o processo de anotação que poderá utilizar mais conhecimento para escolha de um melhor recurso e, possivelmente, com isso melhorar a qualidade desse processo (SILVA, 2010).

1.1 CARACTERIZAÇÃO DO PROBLEMA

Como foi explicado anteriormente, as ontologias são muito úteis para a representação de um domínio, auxiliando no seu entendimento. Entretanto, diversas ontologias sofrem com problemas de omissão de informações relevantes, prejudicando o conhecimento expresso nelas e possivelmente a qualidade dos processos e pesquisas que as utilizam. As ontologias biomédicas não fogem a essa "regra" e, embora sejam amplamente utilizadas no domínio biomédico, contribuindo com diversas pesquisas, ainda apresentam problemas na forma como são construídas e estruturadas. Esse fato tem sido apontado em diversos trabalhos (SMITH;

WILLIAMS; SCHULZE-KREMER, 2003; WROE *et al.*, 2003; SMITH; KÖHLER; KUMAR, 2004; SMITH; KUMAR, 2004; SMITH; ROSSE, 2004; OGREN *et al.*, 2004; OGREN; COHEN; HUNTER, 2005; SMITH *et al.*, 2005; KÖHLER *et al.* 2006; SCHULZ; KUMAR; BITTNER, 2006; GU *et al.*, 2009) e confirmado no decorrer da pesquisa dessa dissertação. Entre os problemas mais graves encontrados nessas ontologias podem-se citar os problemas relacionados à hierarquia (como, por exemplo, deficiência na forma como os conceitos são estruturados e a omissão de termos importantes na hierarquia) e os problemas nas relações (como, por exemplo, relações empregadas de forma equivocada e a omissão de relações aparentemente óbvias).

Grande parte desses problemas interfere diretamente na qualidade das ontologias, prejudicando, por exemplo, o reúso, e em alguns casos mais graves, levando a erros nas pesquisas e nos processos que utilizam essas ontologias (como, por exemplo, a anotação) (SMITH *et.al*, 2005; KÖHLER *et al.*, 2006; SMITH; KÖHLER; KUMAR, 2004).

Por outro lado, conforme abordado em diversos trabalhos (DAHLBERG, 1978; 1981; MICHAEL; MEJINO-JR; ROSSE, 2001; KÖHLER *et al.*, 2006), o campo definição possui um papel importante para o entendimento dos termos que formam a ontologia, pois é através desse campo que é possível contextualizar o termo, explicitando conceitos relacionados e permitindo um melhor entendimento por seres humanos e em alguns casos fornecendo informação para inferência por máquinas. Considere o exemplo a seguir:

Termo: *acetylcholine catabolic process in synaptic cleft* (GO: 0001507)[1]

Definição: *The chemical reactions and pathways resulting in the breakdown of acetylcholine that occurs in the synaptic cleft during synaptic transmission.*

No exemplo mostrado é possível, através da definição, entender melhor o termo *acetylcholine catabolic process in synaptic cleft*, pois a definição o relaciona ao conceito de uma reação química, mostrando seu contexto e ainda descrevendo onde esse conceito ocorre.

Além disso, sabe-se também que, nas ontologias biomédicas, a nomenclatura dos termos e o campo definição contêm informações implícitas que poderiam ser usadas para enriquecer o conhecimento contido nelas. Estas informações implícitas poderiam ser usadas para resolver alguns dos problemas dessas ontologias, como, por exemplo, explicitar relações hierárquicas expressas através da nomenclatura dos termos ou do campo definição e omissas

[1] Nessa dissertação optou-se por deixar os exemplos na língua inglesa com a finalidade de facilitar o entendimento, uma vez que a maior parte das ontologias do domínio biomédico utiliza esse idioma como base.

nas ontologias, as quais poderiam vir a melhorar o processo de reúso e anotação. O exemplo a seguir evidencia esta questão. Nele, é possível extrair, a partir da definição, dois tipos de relações: *is_a* e *occurs_in*.

Termo: *female meiosis I* (GO: 0007144)

Definição: *The cell cycle process whereby the first meiotic division occurs in the female germline.*

Relações implícitas: *female meiosis I* **is_a** *cell cycle process;*

female meiosis I **occurs in** *the female germline.*

1.2 OBJETIVO

O objetivo desse trabalho é utilizar a nomenclatura dos termos e o campo definição das ontologias biomédicas para identificar novas relações e novos termos, com a finalidade de promover uma possível melhora nos processos relacionados ao reúso das ontologias biomédicas.

Acredita-se que um enriquecimento do conhecimento contido nas ontologias através do acréscimo de novas relações e termos implícitos possivelmente traria benefícios para a ontologia, tais como aumento da expressividade e capacidade de inferência, o que poderia refletir diretamente na qualidade e nos processos relacionados ao reúso dessas ontologias. No entanto, isso deve ser feito com cautela, pois poderia também trazer malefícios, como aumento da complexidade da ontologia devido ao acréscimo de novos termos e relações desnecessários para o domínio.

O foco dessa dissertação consiste em investigar a existência de conhecimento implícito e avaliar o quão benéfico esse conhecimento poderia ser para uso das ontologias biomédicas.

1.3 ENFOQUE DE SOLUÇÃO

Para alcançar seu objetivo essa dissertação propõe uma abordagem para extração de informação em ontologias explorando particularmente o domínio biomédico. Embora inicialmente essa abordagem seja elaborada para a extração de termos e relações hierárquicas implícitas, no seu desenvolvimento apontam-se caminhos para que possa ser generalizada

para outras relações e outros tipos de informações através de poucas adaptações.

A abordagem proposta nesse trabalho inclui duas macroetapas, uma responsável por conhecer a ontologia (seu processo de formação e a composição dos seus termos e do campo definição) e a outra responsável pela extração de informação propriamente dita. Ao propor essa divisão, acredita-se que é possível se obter um ganho no processo de extração de informação, uma vez que a aplicação da primeira etapa facilita a exploração do conhecimento que ocorre na segunda etapa, por permitir conhecimento sobre a estrutura de formação do corpus.

Na aplicação dessas duas macroetapas serão utilizados quatro enfoques para auxiliar no processo de extração de informação:

- O **enfoque baseado em dicionário** utiliza recursos existentes na terminologia para extrair informação do corpus. (ANANIADOU; McNAUGHT, 2006).
- O **enfoque baseado em regras** utiliza diversos padrões de formação para extrair informação do corpus. (ANANIADOU; McNAUGHT, 2006).
- O **enfoque baseado em aprendizagem de máquina** utiliza técnicas de aprendizagem de máquina para extração de informação do corpus. (ANANIADOU; McNAUGHT, 2006).
- O **enfoque baseado em estatística** utiliza técnicas estatísticas que se baseiam em diferentes distribuições das colocações do texto no corpus para extração de informação (ANANIADOU; McNAUGHT, 2006).

Ao final da construção e apresentação da abordagem são avaliados os benefícios do uso da abordagem e das novas relações nas ontologias biomédicas, utilizando para isso um determinado cenário como base de aplicação.

1.4 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em sete capítulos, incluindo este de introdução. Os demais capítulos dessa dissertação estão estruturados da seguinte forma.

O Capítulo 2 trata dos conceitos relacionados às ontologias e seus usos na área biomédica, descrevendo suas principais características e também os processos de anotação e

reúso.

O Capítulo 3 faz um estudo sobre o campo definição e também sobre as relações nas terminologias, culminando na discussão da relação hierárquica *is_a*, a qual é foco de estudo desse trabalho.

O Capítulo 4 faz uma revisão sobre extração de informação, mostrando sua aplicabilidade na área de Bioinformática. Além disso, esse capítulo também descreve alguns trabalhos relacionados com a abordagem proposta nessa dissertação.

O Capítulo 5 descreve a abordagem proposta, detalhando as macroetapas, etapas e subetapas necessárias para sua aplicação. Nesse capítulo, também são destacadas as duas principais macroetapas da abordagem. A primeira consiste em conhecer o corpus a ser estudado e a segunda consiste na extração de informação propriamente dita.

O Capítulo 6 apresenta a avaliação experimental realizada nessa dissertação, a qual foi dividida em duas fases. A primeira fase tem por objetivo verificar a qualidade das informações extraídas pela abordagem. A segunda fase tem por objetivo verificar a utilidade e benefícios dessas informações. Além disso, esse capítulo também apresenta alguns detalhes da implementação da abordagem no mecanismo EI-ONTO.

Por fim, o Capítulo 7 contém as considerações finais da dissertação, evidenciando as principais contribuições, limitações e também possíveis trabalhos futuros.

CAPÍTULO 2: ONTOLOGIAS E SEU USO EM BIOINFORMÁTICA

Este capítulo apresenta uma visão geral dos conceitos relacionados às ontologias, como origem, áreas de estudo e aplicações. Depois, restringe seu escopo às ontologias biomédicas descrevendo conceitos e características relacionadas a essas ontologias. Nesse capítulo também são descritos a ontologia Gene Ontology, o consórcio OBO, aspectos de qualidade relacionados às ontologias biomédicas e conceitos de reúso e anotação.

2.1 O CONCEITO DE ONTOLOGIA

Representar o conhecimento sobre um determinado domínio permitindo a inferência por máquinas e o entendimento por seres humanos pode ser extremamente complexo. Devido à capacidade das ontologias poderem representar o conhecimento de certo domínio de maneira formal e explícita, elas têm sido utilizadas em diversas áreas, tais como Engenharia de Software (NARDI; FALBO, 2006), Banco de Dados (DEGGAU; FILETO, 2009), Web Semântica (PICKLER, 2007), Gestão do Conhecimento (GAVA; MENEZES, 2003) e muitas outras.

O termo ontologia é oriundo do grego "ontos" (ser) e "logos" (palavra). Originalmente ontologia representa a palavra aristotélica "categoria", ou seja, o que pode ser utilizado para classificar algo (ALMEIDA; BAX, 2003). Atualmente o termo ontologia é utilizado em diversos campos de estudo, possuindo inúmeros significados não coincidentes, não havendo, pelo menos em princípio, um consenso sobre o que realmente é uma ontologia e em que contexto ela deve ser utilizada (GUARINO; GIARETTA, 1995). Brewster *et al.* (2005), por exemplo, afirmam que as ontologias estão a ponto de tornar-se uma "*quimera*[2]" devido à forma que têm sido utilizadas, podendo servir em qualquer contexto e para qualquer finalidade, possuindo uma infinidade de utilidades e descrições.

Na verdade, essas definições variam de acordo com o campo no qual a ontologia está sendo usada e na forma como está sendo aplicada. Guarino (1998a) explica que existem dois tipos de ontologia. A ontologia com "O" maiúsculo, que se refere a um campo de estudo da Filosofia e a ontologia com "o" minúsculo, que se refere à ontologia aristotélica, tendo sido estudada e reconhecida em diversas áreas de pesquisa, tais como: Ciência da Computação,

[2] Criação absurda da imaginação; fantasia, utopia, sonho.
Fonte: Dicionário Michaelis Online. <http://michaelis.uol.com.br>

Biologia e Ciência da Informação, entre outras áreas. Smith (2010, p.22) corrobora com essa divisão, afirmando que:

> *"O filósofo-ontologista, pelo menos em princípio, tem somente um objetivo: estabelecer a verdade sobre o domínio em questão. No mundo de sistemas de informação, em contraste, uma ontologia é um artefato de software (ou linguagem formal) projetado com um conjunto específico de usos e ambientes computacionais em mente, muitas vezes encomendados por um determinado cliente ou para um programa (aplicativo) em um determinado cenário."*

O Dicionário do Aurélio Online (2010) atribui à Ontologia a descrição filosófica, descrevendo-a como: *"ciência do ser em geral, que considera o ser em si mesmo, independentemente do modo pelo qual se manifesta."*. Nessa mesma linha de pensamento, o guia de termos e conceitos de Epistemologia e Filosofia *The Epistemological Lifeboat* define ontologia como sendo usada para: "*referir à investigação filosófica da existência"* (HJORLAND; NICOLAISEN, 2010). Em Filosofia, ontologia é a disciplina que estuda a natureza e a existência das coisas, estudando todos os aspectos de sua realidade (GLOCK, 2002; WELTY; GUARINO, 2001).

Em outras áreas (como por exemplo, Ciência da Computação, Biomedicina e Ciência da Informação) a definição mais clássica de ontologia é a definição de Gruber (1993, p.199) que descreve ontologia como uma: *"especificação explícita de uma conceitualização"*. Por **conceitualização** entende-se um modelo "abstrato" de um contexto, no caso o domínio o qual a ontologia modela. Essa conceitualização é **explícita** por considerar que termos, definições, propriedades, instâncias e relações utilizados em uma ontologia devem ser definidos explicitamente (DING; FOO, 2002). Gruber (1993, p.199-200) completa essa definição explicando mais detalhadamente o conceito de ontologia:

> *"O termo é emprestado da Filosofia, onde uma ontologia é um relato sistemático da Existência. Para sistemas baseados em conhecimento, o que "existe" é exatamente o que pode ser representado. Quando o conhecimento de um domínio é representado em um formalismo declarativo, o conjunto de objetos que podem ser representados é chamado de universo do discurso. Este conjunto de objetos, e as relações que são descritas entre eles, são refletidas no vocabulário de representação com o qual um "programa" baseado em conhecimento representa o conhecimento. Assim, podemos descrever a ontologia de um "programa", definindo um conjunto de termos representativos. Em tal ontologia, definições associam nomes de entidades no universo do discurso (por exemplo, classes, relações, funções ou outros objetos), com texto legível descrevendo o que os nomes são destinados a denotar, e axiomas formais que restringem a interpretação e uso bem-formado desses nomes (termos)."*

Na Ciência da Computação, ontologias são utilizadas para representação formal de domínios de conhecimento, tendo surgido por volta da década de 90, no âmbito da área de Inteligência Artificial (FENSEL *et al.*, 2001; GRUBER, 2008). Guarino (1998a, p.4) descreve ontologias na Ciência da Computação como: *"um artefato de engenharia, constituído por um vocabulário específico usado para descrever certa realidade, mais um conjunto de pressupostos explícitos sobre o significado pretendido das palavras do vocabulário"*.

Na Ciência da Informação, assim como na Ciência da Computação, ontologias são utilizadas para modelar o conhecimento de um domínio, conforme descrito por Gruber (2008, p.1): *"Em Ciência da Computação e Ciência da informação, ontologia é um termo técnico que denota um artefato que é projetado para uma finalidade, que é a de permitir a modelagem de conhecimento sobre algum domínio, real ou imaginário."*

Nesse trabalho serão estudadas as ontologias do domínio biomédico, um campo que requer fundamentos de muitos outros como, por exemplo, de Ciência da Computação, Ciência da Informação e Biologia. As ontologias biomédicas têm como objetivo principal a representação dos conceitos desse domínio. Como explicitado por Bodenreider e Burgun (2005, p.213):

> *"O propósito da ontologia biomédica é estudar as classes de entidades (ou seja, substâncias, qualidades e processos) que possuem relevância na área biomédica. Exemplos de tais classes incluem substâncias como a válvula mitral e a glicose, qualidades tais como o diâmetro do ventrículo esquerdo e da função catalítica de enzimas, e processos, como a circulação do sangue e a secreção de hormônios. Ao contrário da terminologia biomédica, que congrega os nomes das entidades empregadas na área biomédica, a ontologia biomédica está preocupada com a definição de princípios de classes biológicas e das relações entre elas. Na prática, como são mais que listas de termos, mas não necessariamente cumprem os requisitos de uma organização formal, muitos produtos desenvolvidos por terminólogos e ontologistas biomédicos estão entre terminologias e ontologias e constituem um 'gradiente de ontologia'"*.

2.2 CARACTERÍSTICAS DAS ONTOLOGIAS

Hwang (1999) descreve cinco características importantes para uma ontologia:

1. **Aberta e Dinâmica:** uma ontologia deve possuir seus algoritmos e estruturas abertos e dinâmicos adaptando-se as mudanças e novos desenvolvimentos com o mínimo de ajuda humana.

2. **Escalável e Interoperável:** uma ontologia deve ser escalável e interoperável para facilitar a ampliação do domínio, a adaptação a novos requisitos e a integração com outras ontologias.

3. **Fácil Manutenção:** uma ontologia deve permitir a fácil manutenção de conteúdo a fim de que possa estar sempre atualizada. Para isso a ontologia deve possuir uma estrutura simples e de fácil entendimento por seres humanos.

4. **Consistente Semanticamente:** uma ontologia deve garantir a coerência do domínio representado através de uma semântica estrutural dos relacionamentos entre os conceitos.

5. **Independente do Contexto:** uma ontologia deve ter conceitos representativos, não sendo "vagos" ou "específicos" demais, visando sempre à integração com outras ontologias.

Embora as ontologias nem sempre apresentem um mesmo metamodelo conceitual, existem componentes presentes em grande parte delas. Esses componentes de uma ontologia podem ser definidos como (NOY; McGUINNESS, 2001; ALMEIDA; BAX, 2003; GRUBER, 2008):

- **Classes/Termos:** representam os conceitos de um determinado domínio. As classes podem possuir subclasses ou classes filhas (termos filhos) (utilizadas para representar conceitos mais específicos) e superclasses ou classes pais (termos pais) (utilizadas para representar conceitos mais genéricos).

- **Relações:** representam as interconexões (ligações) entre classes ou instâncias do domínio.

- **Axiomas:** representam restrições (limitações) para uma dada classe (termo), relação ou instância.

- **Instâncias:** representam "casos particulares" (exemplos) das classes no domínio.

Sales, Campos e Gomes (2008), simplificam esse conjunto afirmando que de forma geral, as ontologias são constituídas por:

- **Termos:** componentes que representam os conceitos de um domínio.
- **Relações:** interconexão (ligações) entre os conceitos.
- **Definições:** descrição mais detalhada das características dos conceitos.

A definição de Sales, Campos e Gomes (2008) descreve perfeitamente a estrutura das ontologias biomédicas estudadas nessa dissertação. Por isso, sempre que forem referidos os componentes das ontologias, nessa dissertação, será utilizada essa definição como base.

Quanto à sua classificação, as ontologias podem ser classificadas sobre diferentes aspectos como, por exemplo, quanto ao conteúdo, estrutura, função e aplicação (ALMEIDA; BAX, 2003). Todavia, para o contexto desse trabalho só interessa descrever as ontologias em relação ao grau de formalismo e nível de especificidade (generalidade).

Quanto ao grau de formalismo as ontologias podem ser basicamente classificadas em leves (*lightweight)* e pesadas (*heavyweight*). Ambos os tipos de ontologias possuem conceitos, estrutura taxonômica dos conceitos, relações e propriedades que os descrevem. Entretanto, nas ontologias pesadas adicionam-se regras com o objetivo de permitir um maior poder de inferência, através da delimitação dos conceitos (CORCHO; FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ, 2003). Outros autores (USCHOLD, 1996; USCHOLD; GRUNINGER, 1996; ALMEIDA; BAX, 2003) dividem as ontologias quando ao grau de formalismo em quatro categorias:

- **Altamente informais:** ontologias construídas em linguagem natural (humana), sem qualquer vocabulário de apoio.
- **Informais estruturadas:** ontologias construídas em linguagem natural, porém utilizando-se de um vocabulário controlado (conjunto de palavras restrito), reduzindo, com isso, as ambiguidades.
- **Semiformais:** ontologias construídas em uma linguagem artificial, definida formalmente.

- **Rigorosamente formais:** ontologias construídas com uma semântica meticulosamente formal. A definição dos termos, por exemplo, é feita utilizando uma semântica extremamente formal, teoremas e provas complenitude.

As ontologias altamente informais e informais estruturadas podem ser consideradas como ontologias leves (*lightweight*), e as semiformais e rigorosamente formais como ontologias pesadas (*heavyweight*).

Quanto ao seu nível de especificidade (generalidade) as ontologias podem ser classificadas da seguinte forma (GUARINO, 1997; van HEIJST; SCHREIBER; WIELINGA, 1997; HAAV; LUBI, 2001; ALMEIDA; BAX, 2003):

- **Ontologia de alto nível, de fundamentação, genéricas ou de topo:** descrevem conceitos gerais como ação, coisas, evento, espaço, tempo, matéria, modo, objeto, etc. Essas ontologias são independentes de um domínio ou problema particular.

- **Ontologia de domínio:** descrevem o vocabulário relacionado a algum domínio genérico. Utilizando como exemplo o domínio biomédico poderíamos ter conceitos como aminoácidos, enzimas, genes e proteínas.

- **Ontologia de tarefa:** descrevem o vocabulário relacionado a uma atividade ou uma tarefa genérica. Como, por exemplo, essas ontologias poderiam descrever atividades como anotação, expansão de consultas ou reúso.

- **Ontologias de aplicação:** descrevem conceitos que são especializados nas ontologias de domínio e tarefas. Esses conceitos podem ser exemplificados pelos os papéis desempenhados por entidades de domínio ao executar uma atividade ou tarefa.

A Figura 1 proposta por Guarino (1997) ilustra o nível de dependência entre essas ontologias. As setas indicam as relações de especialização.

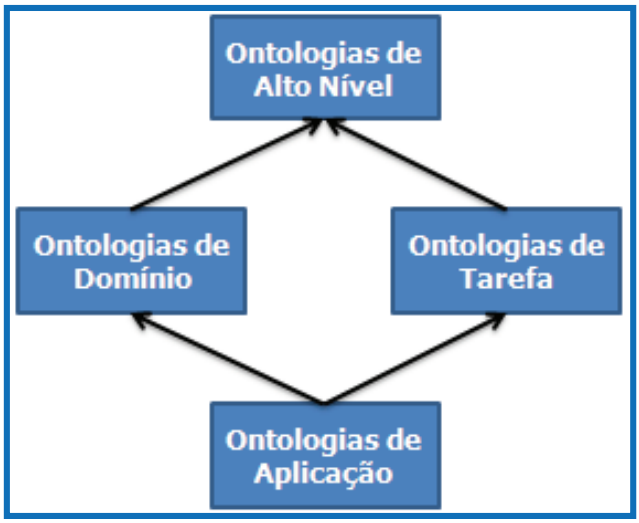

Figura 1: Nível de dependência entre as ontologias
Fonte: Guarino (1997)

Gómez-Pérez e Benjamins (1999) subdividem as ontologias quanto a seu nível de especificidade (generalidade) em mais categorias, porém de maneira equivalente. Além das ontologias de alto nível, domínio, tarefa e aplicação são acrescentadas por Gómez-Pérez e Benjamins (1999) as seguintes categorias:

- **Ontologias de representação do conhecimento:** descrevem as primitivas de representação do conhecimento para que sejam utilizadas para formalizar o conhecimento em paradigmas de representação.

- **Meta Ontologias (Core Ontologies):** descrevem conceitos mais genéricos para que esse conhecimento possa ser reutilizável em outros domínios. Esse tipo de ontologias apresenta um metamodelo que geralmente é utilizado como base para a construção de ontologias mais específicas em um domínio complexo.

- **Ontologias linguísticas:** descrevem conceitos de uma determinada linguagem. O exemplo mais clássico desse tipo de ontologia é o WordNet (MILLER, 1995).

- **Ontologias de tarefas de domínio:** descrevem conceitos característicos das ontologias de tarefas, porém diferentemente dessas ontologias, os conceitos são reutilizáveis somente dentro de determinados domínios e não através de domínios como ocorre nas ontologias de tarefa.

- **Ontologias de método:** descrevem formalmente definições de conceitos e relações importantes em um domínio, para que com isso se possa alcançar uma tarefa

específica através da aplicação de um processo de raciocínio ou método específico (GÓMEZ-PÉREZ; BENJAMINS, 1999; CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999).

2.3 ONTOLOGIAS EM BIOINFORMÁTICA E BIOMEDICINA

Bioinformática é a área de conhecimento que surgiu da aplicação de abordagens, métodos, ferramentas e técnicas computacionais as áreas de pesquisa ligadas à Biologia e à Biomedicina, como, por exemplo, a área Genômica. O termo Bioinformática foi proposto por Hwa Lim na década de 80, no entanto ganhou visibilidade somente na década de 90 ao ser relacionado ao campo de pesquisas genômicas (GOODMAN, 2002). Essa área engloba toda a junção de esforços e parcerias entre biologistas e informatas, permitindo agregar às pesquisas ligadas a Biologia e Biomedicina e suas áreas derivadas diversos conhecimentos consolidados da Ciência da Computação como, por exemplo: gerenciamento de informações, *data warehouse* e mineração de dados (LACROIX; CRITCHLOW, 2003).

Um dos grandes desafios da Bioinformática, atualmente, é tratar o grande volume de informações geradas pelas pesquisas ligadas a área biológica e biomédica, principalmente da área Genômica. Para manipulação e busca desse grande volume de informação torna-se imprescindível a utilização de técnicas de mineração de texto e a utilização de mecanismos para representação formal do conhecimento (domínio), como, por exemplo, as ontologias (TARCZY-HORNOCH; MINIE, 2005).

Com o intuito de ajudar na manipulação das informações as ontologias biomédicas têm sido muito utilizadas na área de Bioinformática (BODENREIDER; STEVENS, 2006). O objetivo dessas ontologias é representar formalmente classes de entidades (como, por exemplo, substâncias qualidades e processos) do contexto biomédico, com a finalidade de servir como apoio para diversas atividades, como por exemplo, a anotação de recursos (BODENREIDER; BURGUN, 2005). Exemplos dessas classes incluem (BODENREIDER; BURGUN, 2005):

- **Substâncias:** glicose.
- **Qualidades:** função catalítica das enzimas.

- **Processos:** circulação do sangue**.**

2.3.1 ONTOLOGIAS DO CONSÓRCIO OBO

O consórcio *Open Biomedical Ontologies* (OBO) (2010a) é um esforço colaborativo entre diversos grupos, a exemplo de: *Biomedical Informatics Research Network* (BIRN) (2010), *Berkeley Bioinformatics Open Source Project* (BBOP) (2010), *Immunology Database and Analysis Portal* (ImmPort) (2010), *National Center for Biomedical Ontology* (NCBO) (2010) e *Ontology Research Group* (ORG) (2010), entre outros. Esse consórcio visa à definição de normas e padrões para criação, manutenção e reformulação de ontologias ortogonais e interoperáveis para uso compartilhado em diferentes domínios biomédicos (SMITH *et al.*, 2007; OBO, 2010a).

Além disso, esse consórcio é o mantenedor do mais importante repositório de terminologias da área biomédica, congregando, atualmente, mais de 75 terminologias, entre elas: *BRENDA tissue/enzyme source* (2010)*, Chemical entities of biological interest* (ChEBI) (2010), *Integrating Network Objects with Hierarchies* (INOH) (2010), *Pathway Ontology* (2010) e a *Gene Ontology* (GO) (2010) a principal terminologia do consórcio OBO (SMITH *et al.*, 2007).

Ao se associarem ao consórcio OBO os mantenedores de uma ontologia devem comprometer-se com os princípios disseminados pela OBO para criação e manutenção de ontologias, que são os seguintes (OBO, 2010b):

1. A ontologia deve ser aberta e disponível para ser utilizada por todos, entretanto, a origem da ontologia deve ser sempre reconhecida, não podendo ser alterada e redistribuída com o nome original ou com os mesmos identificadores.

2. A ontologia deve ser expressa em uma sintaxe reconhecida, como, por exemplo, sintaxe OBO ou OWL.

3. A ontologia deve possuir um único identificador dentro do consórcio OBO.

4. A ontologia deve informar a sua versão, diferenciando uma versão mais atual de outras mais antigas.

5. A ontologia deve informar claramente seu conteúdo e temática.

6. A ontologia deve ter todos os seus termos definidos.

7. A ontologia deve seguir o padrão de definição descrito no OBO *Relation Ontology* (OBO, 2010c).

8. A ontologia deve possuir ampla documentação sobre sua temática e escopo.

9. A ontologia deve possuir um bom conjunto de usuários. Para isso a ontologia deve ser formulada para servir em diferentes contextos.

10. A ontologia deve ser construída e mantida de forma colaborativa com os outros membros do consórcio OBO.

2.3.2 GENE ONTOLOGY

A terminologia do consórcio OBO mais utilizada, atualmente, é a *Gene Ontology* (2010a). O projeto da *Gene Ontology* foi iniciado em 1998 com a finalidade de permitir um vocabulário hierárquico padronizado para descrição de componentes celulares, funções moleculares e funções biológicas (LEWIS, 2004; GENE ONTOLOGY CONSORTIUM, 2004). O projeto inicialmente contemplava conceitos de três grupos de banco de dados biológicos: *FlyBase* (2010), *Mouse Genome Informatics* (MGI) (2010) e o *Saccharomyces Genome Database* (SGD) (2010) (LEWIS, 2004). Com o passar dos anos, a GO passou a contemplar muitas outras bases de dados (GENE ONTOLOGY CONSORTIUM, 2004). O projeto da GO tem três objetivos principais, que são listados a seguir (GENE ONTOLOGY CONSORTIUM, 2004):

1. Desenvolver, por meio de uma ontologia, um conjunto de vocabulários controlados para descrição dos domínios principais da Biologia Molecular como, por exemplo, sequências biológicas.

2. Disponibilizar os termos da GO para auxílio no processo de anotação de genes ou produtos de genes em bases biológicas.

3. Fornecer um repositório público central para permitir o acesso universal à

ontologia, aos conjuntos de anotações e aos mecanismos (como, por exemplo, mecanismos de visualização e navegação) desenvolvidos para uso em conjunto com as informações da GO.

A *Gene Ontology* divide-se em três ramos: Componente Celular (*Cellular Component*), Função Molecular (*Molecular Function*) e Processo Biológico (*Biological Process*) (LEWIS, 2004). O ramo Componente Celular descreve os componentes de uma célula. Exemplos desses componentes são as estruturas anatômicas (como, por exemplo, retículo endoplasmático rugoso ou núcleo) ou grupos de produtos de genes (como, por exemplo, ribossomo). O ramo Função Molecular descreve atividades que ocorrem no nível molecular como, por exemplo, atividades catalíticas e atividades de transporte. O ramo Processo Biológico descreve uma série de eventos (objetivos biológicos) executados por um ou mais conjuntos de funções moleculares, como: processo fisiológico celular ou transdução de sinal (GENE ONTOLOGY, 2010b; GENE ONTOLOGY CONSORTIUM, 2004).

A GO está em constante crescimento, sendo atualizada frequentemente. Para facilitar o acesso a esses dados, dezenas de mecanismos vêm sendo desenvolvidos. Um dos mecanismos mais populares é o AmiGO (GENE ONTOLOGY, 2010c). Esse mecanismo permite, através de uma interface gráfica via web, a busca de termos, definições, sinônimos e anotações, fornecendo diversas possibilidades de visualização da hierarquia dos termos (GENE ONTOLOGY CONSORTIUM, 2004).

2.3.3 QUALIDADE DAS ONTOLOGIAS BIOMÉDICAS

Embora amplamente utilizadas pelas pesquisas na área biomédica, grande parte das ontologias desse domínio, encabeçadas pela GO, possuem deficiências na forma como foram construídas e estruturadas, ocasionando diversos problemas, como descrito por Smith, Köhler e Kumar (2004, p.1): "*Princípios formais que regem as melhores práticas de classificação e definição foram, por muito tempo, negligenciados na construção de ontologias biomédicas, resultando em consequências negativas para a integração de dados e para o alinhamento de ontologias.*".

Algumas deficiências encontradas nas ontologias, discutidas na literatura (SMITH;

WILLIAMS; SCHULZE-KREMER, 2003; WROE *et al.*, 2003; SMITH; KÖHLER; KUMAR, 2004; SMITH; KUMAR, 2004; SMITH; ROSSE, 2004; OGREN *et al.*, 2004; OGREN; COHEN; HUNTER, 2005; SMITH *et al.*, 2005; KÖHLER *et al.* 2006; SCHULZ; KUMAR; BITTNER, 2006; GU *et al.*, 2009) e levantadas no decorrer do estudo desse trabalho são:

Problemas na Hierarquia:

- Deficiência na estruturação dos conceitos (conceitos muito específicos ligados diretamente a conceitos muito genéricos).
- Clara omissão de conceitos, aparentemente óbvios, na hierarquia.

Problemas nas Relações:

- Poucas relações disponíveis para expressar o conhecimento de um determinado domínio.
- Relações empregadas de forma equivocada.
- Falta de relacionamento entre conceitos que claramente deveriam estar relacionados.

Problemas na definição dos termos:

- Informações implícitas no campo definição inconsistentes com o conhecimento representado na ontologia.
- Definições em linguagem natural "livre", dificultando a compreensão por máquinas.
- Definições circulares.
- Falta de padronização das definições em algumas ontologias.
- Definições dentro de definições, ou seja, uma definição de um conceito específico, definindo também um segundo conceito em seu corpo.

Problemas Contextuais:

- Falta de documentação formal da ontologia.

- Falta de descritores sobre a temática e escopo da ontologia.
- Falta de padronização para a nomenclatura dos termos das ontologias biomédicas, fazendo com que conceitos com nomes idênticos representem conceitos diferentes e conceitos com nomes diferentes representem um mesmo conceito.
- Falta de iniciativas para reúso das ontologias que algumas vezes possuem conceitos similares ou complementares, porém não se relacionam explicitamente.

Os problemas descritos anteriormente interferem diretamente na qualidade dessas ontologias. Entretanto, técnicas de mineração, processamento e extração de textos podem ser usadas para ajudar a corrigir alguns deles, visando à melhoria da qualidade e da expressividade das ontologias (KÖHLER *et al.*, 2006).

2.3.4 APLICABILIDADE DAS ONTOLOGIAS BIOMÉDICAS

As ontologias possuem diversos usos no domínio biomédico. Pode-se citar como exemplo a anotação de recursos (BELLOZE, 2007), reúso (SILVA, 2010) e expansão de consultas (ELIAS, 2009). O reúso é especialmente importante, nesse contexto, pois viabiliza o uso conjunto de ontologias existentes.

Nesse trabalho, somente serão descritos com mais detalhes os conceitos de anotação e alinhamento (ligados diretamente ao reúso das ontologias), por esses estarem relacionados diretamente com a pesquisa desenvolvida.

2.3.4.1 ANOTAÇÃO

O processo de anotação de recursos genômicos consiste em mapear características ou papéis biológicos às sequências de genes (FRISHMAN; VALENCIA, 2008). Existem dois tipos de anotação: as anotações externas, que são as anotações armazenadas e manipuladas em fontes de dados públicos e as internas, que são as anotações armazenadas e manipuladas em fontes de dados no ambiente de pesquisa (LEMOS, 2004; ALMEIDA, 2006; GUIMARÃES;

CAVALCANTI, 2009). As anotações internas, por sua vez, dividem-se em (LEMOS, 2004; ALMEIDA, 2006; GUIMARÃES; CAVALCANTI, 2009):

- **Manuais:** anotações feitas manualmente pelo anotador.
- **Automáticas:** anotações feitas automaticamente por mecanismos de análise.
- **Semiautomáticas:** anotações feitas automaticamente, por mecanismos de análise, acrescidas das informações observadas pelo anotador.
- **Importadas:** anotações são obtidas (importadas) através de fontes de dados públicas como, por exemplo, o GENBANK (BENSON *et al.*, 2004).

O processo de anotação de recursos genômicos geralmente ocorre em três etapas. Na primeira etapa os dados são importados a partir de fontes de dados públicas. Nessa etapa, geralmente, são utilizados diversos mecanismos e ferramentas para realizar as anotações automaticamente (como, por exemplo, BLAST (ALTSCHUL *et al.*, 1990)). Na segunda etapa é realizada a identificação das sequências de genes, utilizando critérios pré-estabelecidos, através dos dados obtidos na primeira etapa. Por fim, na terceira etapa, ocorre a anotação manual feita pelo anotador (LEMOS; SEIBEL; CASANOVA, 2003; BELLOZE, 2007).

Atualmente existem diversos mecanismos e ferramentas voltados para anotação de recursos, como por exemplo: Apollo (LEWIS *et al.*, 2002) e BioNotes (LEMOS; SEIBEL; CASANOVA, 2003). Para ajudar o trabalho do anotador geralmente as ferramentas de anotação de recursos genômicos devem (LEMOS; SEIBEL; CASANOVA, 2003):

- Auxiliar os pesquisadores a explorarem, testarem e avaliarem com maior precisão suas hipóteses.
- Acessar as anotações de dados públicas e fornecer mecanismos de comparação com resultados da pesquisa.
- Executar programas de análise para criação e validação das anotações automáticas.
- Fornecer interface apropriada para análise das anotações existentes e para a criação manual de novas anotações.

No processo de anotação genômica a utilização de ontologias é de grande importância,

uma vez que as ontologias fornecem um vocabulário controlado para a anotação de recursos, permitindo a identificação e um entendimento global do contexto do termo anotado, independente do grupo que realizou a anotação (SILVA, 2010). A ontologia mais utilizada nesse processo é a *Gene Ontology* (2010a). Essa ontologia conta com diversos mecanismos de apoio (como, por exemplo, BLAST (ALTSCHUL *et al.*, 1990) e AmiGO (GENE ONTOLOGY, 2010c)) para auxílio na análise e busca de sequências genômicas.

Outra ontologia muito utilizada para auxiliar a anotação de recursos genômicos é a Sequence Ontology. Essa ontologia descreve, através de seus termos, características biológicas que podem ser mapeadas nas sequências de genes analisadas (EILBECK *et al.,* 2005).

2.3.4.2 REÚSO DE ONTOLOGIAS

O processo de reúso de ontologias consiste em integrar o conhecimento contido em duas ou mais ontologias objetivando um ganho de conhecimento sobre um determinado domínio (CAMPOS, 2009a). O reúso de ontologias ocorre principalmente através do mapeamento (*mapping*) de ontologias. Esse processo pode ser muito complicado, principalmente, devido a não padronização de desenvolvimento/criação das ontologias, o que faz com que cada desenvolvedor utilize um padrão particular, e à utilização de linguagem natural, que dificulta a inferência de informações por máquinas. Conforme descrito por Maedche e Staab (2002, p.251): *"A desvantagem seria que todas as ontologias do mundo real que conhecemos não apenas especificam sua conceitualização por estruturas lógicas, mas em grande medida, também por referência a termos que são explicitados através do uso da linguagem natural."*.

Para auxiliar no mapeamento de ontologias existem diversos mecanismos como, por exemplo, Prompt (NOY; MUSEN, 2000) e FOAM (EHRIG; SURE, 2005). Segundo Noy (2004), os mecanismos com esse propósito devem analisar os seguintes recursos:

- Nomes dos conceitos.
- Hierarquia das classes (relações de subclasse e superclasse).
- Definições de propriedades (domínios, intervalos, restrições).

- Instâncias das classes.
- Definições (descrições) das classes.

Os tipos mais comuns de mapeamento entre ontologias são: o alinhamento (Figura 2) e a união (integração ou junção) (Figura 3) (CAMPOS *et al.*, 2009). O alinhamento entre ontologias consiste estabelecer vínculos (ligações) entre conceitos equivalentes de duas ontologias (SILVA, 2010). O alinhamento ocorre baseado em diversas medidas de similaridade entre os conceitos da ontologia, como: similaridade de *label* e taxonômica (similaridade dos conceitos na hierarquia) (SILVA, 2010). A diferença entre o alinhamento de ontologias e a união (integração ou junção) está no resultado gerado, pois enquanto que na união uma nova ontologia é construída, como sendo a combinação das anteriores, o alinhamento mantém as ontologias inalteradas em seus locais de origem, gerando vínculos entre os conceitos equivalentes das ontologias alinhadas (CAMPOS *et al.*, 2009).

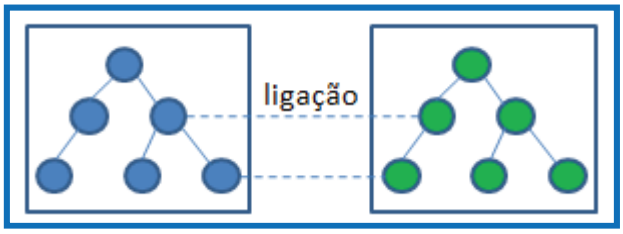

Figura 2: Alinhamento de Ontologias

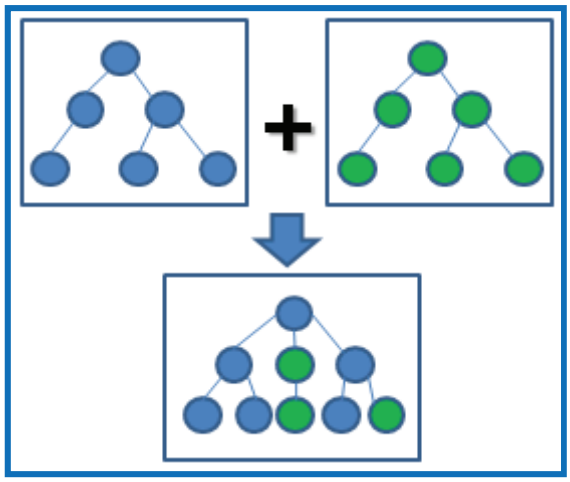

Figura 3: União (Integração ou Junção) de Ontologias

A maior parte dos alinhamentos entre ontologias biomédicas é feita utilizando a ontologia *Gene Ontology*. Embora essa ontologia cubra grande parte das características que os biólogos utilizam para anotação de recursos, a utilização do alinhamento de ontologias permite um enriquecimento da pesquisa através da utilização de outras ontologias mais específicas em conjunto com a GO (SILVA, 2010). O Alinhamento permite, entre outras

utilidades, que o anotador escolha um melhor conceito, baseando-se na comparação da hierarquia em mais de uma ontologia. Isso talvez não fosse possível sem o alinhamento, uma vez que a utilização de duas ontologias para anotação geraria um trabalho adicional para o anotador que teria de percorrer cada uma das ontologias, individualmente, em busca do termo que seria utilizado para a anotação (SILVA, 2010).

2.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma discussão sobre ontologias, mostrando alguns conceitos relacionados importantes para o entendimento dessa dissertação. Além disso, também fez um pequeno resumo sobre a qualidade e utilização das ontologias biomédicas, foco de pesquisa desse trabalho.

Na área biomédica as ontologias, encabeçadas pela GO, vêm sendo utilizadas com muita frequência como um apoio a diversas pesquisas. Entretanto essas ontologias ainda possuem diversos problemas que interferem diretamente em sua qualidade. Contudo visando um aumento da expressividade do conhecimento, qualidade e correções de erros, existem diversas iniciativas, sendo o consórcio OBO a mais importante delas.

O próximo capítulo descreve definições e as relações nas terminologias, entre elas as da área biomédica, dando ênfase principalmente às relações hierárquicas (*is_a*), as quais são objeto de pesquisa neste trabalho.

CAPÍTULO 3: O CAMPO DEFINIÇÃO E AS RELAÇÕES NAS TERMINOLOGIAS

Este capítulo apresenta uma discussão sobre o campo definição evidenciando sua utilidade, importância e diretrizes de construção e manutenção. Além disso, discute também as relações nas ontologias e terminologias de uma forma geral. Ao final desse capítulo são levantadas questões relativas à aplicação desses dois conceitos às ontologias biomédicas, como a importância, o papel e os principais tipos de definições e relações.

3.1 CAMPO DEFINIÇÃO

O campo definição, estudado amplamente nesse trabalho, tem suma importância nas ontologias e em outras terminologias, pois é através desse campo que se é possível complementar o conhecimento expresso nessas terminologias, através da associação com conceitos conhecidos (DAHLBERG, 1981). Diversos trabalhos (DAHLBERG, 1978; 1981; MICHAEL; MEJINO-JR; ROSSE, 2001; KÖHLER *et al.*, 2006) abordam a importância de uma definição formal e estruturada, permitindo um melhor entendimento por seres humanos e a inferência de conhecimento feita por máquinas.

Neste contexto, problemas na definição de conceitos têm sido objeto de diversos estudos ao longo dos anos. Em 1982, por exemplo, o *Groupe Interdisciplinaire de Recherche Scientifique et Appliquée en Terminologie* (GIRSTERM) já discutia as especificidades da definição em terminologia (por exemplo: O quê definir? Como definir? Por que definir? A sinonímia dos termos se compara à sinonímia das palavras?) através de um Colóquio Internacional de Terminologia sob o tema "Problemas da definição e da sinonímia em terminologia" (CAMPOS, 1994; GOMES; CAMPOS, 1994).

Nessa linha de pensamento, Dahlberg (1978, p.149) apresenta a definição como sendo: *"o estabelecimento de uma equivalência entre o termo (o definiendum) e as características necessarias do referente de um conceito (o definiens) com a finalidade de delimitar o uso do termo no discurso"*. Posteriormente, Dahlberg (1981, p.17) reescreve e expande essa definição, porém de maneira análoga à apresentada anteriormente, caracterizando esse conceito como: *"uma equivalência entre o definiendum (o que deve ser definido?) e o definiens (como algo deve ser definido?) com a finalidade de delimitar o entendimento do definiendum em qualquer caso de comunicação"*

3.1.1 TIPOS DE DEFINIÇÃO

Dahlberg (1978; 1981) descreve três tipos de definições: definição nominal, ostensiva e conceitual (real). Para descrição de cada tipo de definição será usado o Triângulo do Conceito de Dahlberg (1978) (Figura 4) que descreve os três elementos que formam um conceito: o referente, as características e o termo.

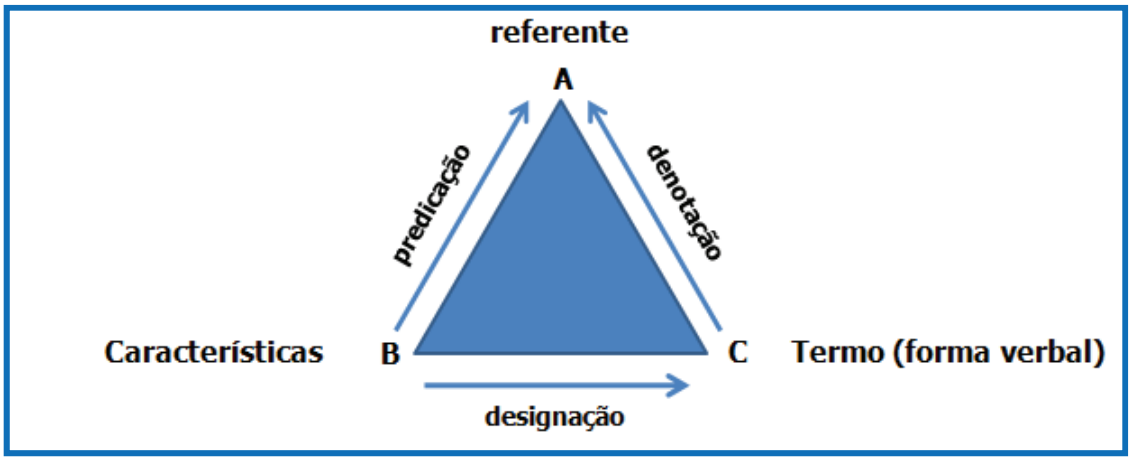

Figura 4: Triângulo do Conceito de Dahlberg
Fonte: DAHLBERG (1978, p.144)

A **definição nominal** é a definição onde o *definiendum* é uma expressão verbal (o termo) e o *definiens* é uma equivalência textual do termo (expressão das características), ou seja, o termo é igualado à expressão das características negligenciando o referente (C (termo) = B (características) negligenciando A (referente)). Esse tipo de definição é geralmente encontrado nos dicionários (DAHLBERG, 1978; 1981; CAMPOS, 2007).

A **definição ostensiva** é a definição onde *definiens* é estabelecido por uma indicação para o referente nomeado pelo *definiendum*, ou seja, o termo é igualado ao referente negligenciando as características (C (termo) = A (referente) negligenciando B (características)). Na definição ostensiva o entendimento e a interpretação dependem da percepção do observador (DAHLBERG, 1978; 1981; CAMPOS, 2007).

A **definição conceitual ou real** é a definição que contém os três elementos do conceito: o referente, as características e o termo. Essa definição ocorre quando o *definiens* contém as características necessárias de um referente nomeadas pelo *definiendum,* ou seja, o termo é igualado às características do referente (C = B com referência A) (DAHLBERG, 1978; 1981; CAMPOS, 2007).

Nas ontologias biomédicas só são encontradas as definições nominais e conceituais (reais). Devido às definições ostensivas não descreverem um termo de uma forma textual e sim representá-lo de uma forma "realista", "simbólica", "gráfica" ou "icônica", que depende, principalmente, do entendimento e interpretação do observador.

As estruturas de *definiens* são utilizadas para permitir um melhor entendimento dos conceitos que formam o *definiendum*. A estrutura de *definiens* mais conhecida e mais antiga é *genus proximum* e *differentia specifica* (DAHLBERG, 1978). Essa estrutura relaciona um dado conceito a um conceito equivalente ou similar (*genus proximum*) e diferencia (especifica) esse conceito através do acréscimo de características (*differentia specifica*), possibilitando com isso, que o conceito principal seja relacionado com algo conhecido ou próximo ao observador (DAHLBERG, 1978; GOMES; CAMPOS, 1994).

Além desse *definiens,* também pode ser encontrado na literatura o *genus supremum.* Esses *definiens* denotam as características especificadoras tendo os mesmos princípios das relações hierárquicas, uma vez que ele relaciona um dado conceito a um conceito mais geral (*genus supremum*) (DAHLBERG, 1978; GOMES; CAMPOS, 1994). O *definiens genus supremum* geralmente são utilizados para descrever conceitos de uma "categoria" ou "propriedade" ou que são descritos por uma ação (DAHLBERG, 1978; GOMES; CAMPOS, 1994). Nas ontologias biomédicas o *genus supremum* é verificado com mais frequência que o *genus proximum*, pois a maior parte das definições dessas ontologias associa o conceito que está sendo definido a um conceito hierarquicamente superior como, por exemplo, em: "um ciclo de processo...", "qualquer processo..." ou "uma reação química...", onde, claramente, verifica-se o emprego desse tipo de estrutura.

Ao analisar as definições sobre a perspectiva de suas estruturas de formação é possível dividi-las em três tipos: genérica, partitiva e funcional (DAHLBERG, 1981).

- A genérica permite identificar a categoria do conceito expresso, utilizando-se de princípios de *genus proximum* e *differentia specifica*. A primeira característica (conceito) da definição é o *genus proximum* e as demais são as *differentia specifica* (DAHLBERG, 1981; CAMPOS, 2010).

  Ex.: Congregação = df uma sociedade de membros da Igreja Católica Romana para fins religiosos ou beneficentes[3].

[3] Exemplos retirados de Dahlberg (1981).

- ***Genus*:** Congregação.
- ***Differentia specifica (1)*:** membros da Igreja Católica Romana.
- ***Differentia specifica (2)*:** para fins religiosos ou beneficentes.

- A partitiva permite reconhecimento dos componentes (o todo e suas partes) de um conceito, identificando as características e enumerando-as (DAHLBERG, 1981; CAMPOS, 2010).

  Ex.: Sociedade =df um grupo de pessoas que formam uma única comunidade[3].
  - **Todo:** grupo.
  - **Partes:** pessoas, membros do grupo.

- A funcional permite identificar a função/especialização do conceito em uma determinada área, ou seja, essa estrutura identifica o seu referente pelo resultado de alguma operação sofrida por alguma pessoa ou coisa (DAHLBERG, 1981; CAMPOS, 2010).

  Ex.: Crescimento do Produto Nacional = df valor total da produção anual de bens e serviços[3].
  - **Operação:** soma de toda a produção (bens e serviços) produzida por uma nação em certo ano.
  - **Sujeito Lógico:** bens e serviços de uma nação.
  - **Predicado Lógico:** soma dos valores totais da produção.
  - **Complemento (Condição):** ano base.

3.1.2 CARACTERISTICAS DE UMA BOA DEFINIÇÃO

Uma boa definição, segundo Dahlberg (1981), tem como principal objetivo estabelecer uma associação entre algo desconhecido com algo conhecido. Para alcançar esse objetivo, Dahlberg (1981) propõe regras para a forma e para o conteúdo das definições:

- Regras relativas à forma de construção das definições:

  - **Simplicidade:** a definição deve conter somente as características relevantes para entendimento de um determinado referente.

- **Clareza:** a definição deve possuir termos, expressões e palavras com significado claro (de fácil entendimento) e definidos dentro do contexto do domínio.
- **Nível:** a definição deve ser formulada utilizando a linguagem apropriada para seu público alvo e baseando-se no conhecimento desse público sobre o assunto.
- **Justaposição na Definição:** a definição deve conter conceitos equivalentes, sinônimos, ou quase sinônimos, para permitir um melhor entendimento e associação dos conceitos pelo leitor.

- Regras relativas ao conteúdo das definições:

  - **Correspondência com o referente:** a definição deve garantir que o *definiendum* e *definiens* possuam um referente único e comum.
  - **Completude da Definição:** a definição deve possuir todas as características necessárias de um referente de forma estruturada.
  - **Adequação da definição:** a definição deve ser restrita ao seu contexto de utilização. Para isso deve-se tomar cuidado com a inclusão ou remoção de características para não restringir ou expandir o seu uso.
  - **Definição não subjetiva:** a definição deve ser sempre uma verdade para um determinado contexto, não podendo ser subjetiva contendo pontos de vista específicos.
  - **Não misturar conceitos:** a definição deve sempre descrever a real característica do conceito definido e não propor interpretações especiais.
    Ex.: Progresso = desenvolvimento das indústrias de tecnologia.
    Nesse exemplo, não é utilizada a definição real de progresso, mas sim uma interpretação especial para as indústrias de tecnologia.
  - **Não criar definições circulares:** a definição deve ser criada de tal forma que não permita definições circulares em uma terminologia. As definições circulares ocorrem de duas formas:
    1. Com a utilização do *definiens* de uma definição como *definiendum* de outra.
       Ex.:
       Definição – é a descrição de conceitos
       Descrição – é a definição de conceitos

2. Com a utilização *genus proximun* no *definiens,* que foi definido em outra parte do sistema como um subconceito do conceito que está sendo definido.
   Ex.:
   Relação = é uma ligação entre duas classes
   Ligação = é uma relação entre duas classes
   Substituindo a segunda definição pelo termo da primeira é verificada a definição circular: Relação = é uma relação entre duas classes.

Copi e Cohen (2004) *apud* Köhler *et al.* (2006) ratificam algumas regras propostas por Dahlberg (1981), afirmando que uma boa definição deve seguir 5 regras, baseadas nos princípios de definição Aristotélica:

1. Foco nas características principais.
2. Não criar definição circulares.
3. Descrever a correta extensão da definição.
4. Não utilizar linguagem figurativa ou obscura.
5. Ser sempre afirmativa em vez de negativa.

3.2 RELAÇÕES NAS ONTOLOGIAS

As relações são muito importantes em uma ontologia, pois é através delas que os termos são interligados (conectados) promovendo, entre outras coisas, a criação da estrutura hierárquica e também o aumento da expressividade e semântica (GUARINO, 1995). Dahlberg (1978) divide as relações entre conceitos em:

- **Quantitativas** – medem a quantidade e a similaridade das características dos conceitos. Essas relações podem ser dividas em quatro tipos (DAHLBERG, 1978; SALES, 2006):

  - **Identidade Conceitual:** ocorre quando as características existentes em dois conceitos são as mesmas.
  - **Inclusão Conceitual:** ocorre quando todas as características de um dado conceito estão inseridas em grande número de características de outro conceito.

- **Interseção Conceitual:** ocorre quando há uma sobreposição das características de dois conceitos.
- **Disjunção Conceitual:** ocorre quando não há nenhuma similaridade entre as características de dois conceitos.

- **Qualitativas**, que podem ser subdividas em (DAHLBERG, 1978; SALES, 2006):

  - **Formais:** são as relações estabelecidas segundo os tipos dos conceitos, isto é de acordo com os referentes conceituais. O estabelecimento das hierarquias dos conceitos é estruturado através de facetas de acordo com as categorias.
  - **Materiais ou Ontológicas:** são relações estabelecidas segundo a categoria fundamental do objeto de um determinado conceito. Esse tipo de relação pode ser divido em:
    - **Relacionamento Hierárquico**: são as relações entre: gênero/espécie, espécie/espécie e espécie/indivíduos.
    - **Relacionamento Partitivo:** são as relações entre o todo e suas partes, entre partes e também entre partes e subpartes.
    - **Relacionamento de Oposição:** são as relações de contradição, contrários e PNI (Positivo–Neutro–Indiferente).
    - **Relacionamento Funcional:** são as relações entre os componentes de uma declaração/proposição, essas relações dependem da semântica das atividades em que são utilizadas.

Sales (2006; 2007) propõe, a partir de um estudo em sua dissertação de mestrado, a classificação das relações nas ontologias em dois tipos:

- **Relações Categoriais:** relações que revelam duplas de categorias Ex.: causa-efeito.

- **Relações Formais:** relações que revelam tipos de ligação entre duplas de categorias. Ex.: *is_a, occurs_in* e *part_of*.

Ainda segundo Sales (2006; 2007), cada um desses tipos pode ser subdivido em três categorias: relações genéricas, relações partitivas e relações funcionais.

Tabela 1: Sistematização das Relações Genéricas

Relações Categoriais Genéricas	Relações Formais Genéricas
gênero-espécie	has_type
espécie-gênero	is_a
espécie-instância	is_instanced_by
instância-espécie	is_ instance_of

Fonte: Sales (2006; 2007)

Tabela 2: Sistematização das Relações Partitivas

Relações Categoriais Partitivas	Relações Formais Partitivas
objeto-constituinte	has_constituent
constituinte-objeto	has_holo_constituent
objeto-unidade	has_unity
unidade-objeto	has_holo _unity
coleção-elemento	has_element
elemento-coleção	has_collection
massa-porção	has_portion
porção-massa	has_mass
objeto-material	has_material
material-objeto	has_holo_material
área-lugar	has_locate
lugar-área	has_holo_locate
zona-região	has_locate
região-zona	has_holo_locate
caráter-atividade	has_stage
atividade-caráter	has_holo_stage
processo continuo-ato	has_stage
ato-processo contínuo	has_holo_stage
processo descontínuo-fase	has_stage
fase-processo descontínuo	has_holo_stage
grupo-membro	has_member
membro-grupo	has_holo_stage
classe-membro	has_member
membro-classe	has_holo_stage
substância-particular	has_substance
particular-substância	has_holo_substance
conjunto-subconjunto	has_subset
subconjunto-conjunto	has_set
conjunto-elemento	has_element
elemento-conjunto	has_set
todo/sub-parte temporal	has_subpart
sub_parte temporal – todo	has_holo
ciência-objeto de estudo	has_branch
objeto de estudo – ciência	has_hole_branch

Fonte: Sales (2006; 2007)

Tabela 3: Sistematização das Relações Funcionais

Relações Categorias Funcionais	Relações Formais Funcionais
personalidade – personalidade	is_a_product_of
	is_an_instrument_of
	transformation_into
	derives_from
	derivated_of
	interacts_with
	manifestation_of
	analyses
	measures
	creates
	placed_in
	located_in
	adjacent_to
	surrounds
	traverses
processo – processo	method-of
	interacts_with
	is_subevent_of
	is_phase_of
	is_stage_of
	brings_about
	prevents
processo – personalidade	brings_about
	affects
	disrupts
	destroys
	results_in
	placed_in
espaço-espaço	adjacent_to
processo – espaço	occurs_in
medida – personalidade	is_a_measure_of
medida – processo	is_a_measure_of
personalidade – processo	used_in
	method _of
	caused_by
	affects
	performs / carries_out
	used_for
	result_of
	analyzes
	measures
	diagnoses
	has_agent/ has_patient
personalidade – propriedade	has
propriedade – personalidade	is_a_property_of
	is_a_counteragent_of
propriedade-processo	is_a_property_of

Fonte: Sales (2006; 2007)

- **Relações Genéricas:** são as relações "é tipo de". A Tabela 1 ilustra as relações genéricas sistematizadas por Sales (2006; 2007).

- **Relações Partitivas:** são as relações que relacionam o todo e suas partes. A Tabela 2 ilustra as relações partitivas sistematizadas por Sales (2006; 2007).

- **Relações Funcionais:** são as relações que descrevem a associação de um dado objeto com o mundo ou com a função de um objeto em determinado contexto. A Tabela 3 ilustra as relações funcionais sistematizadas por Sales (2006; 2007).

3.3 DEFINIÇÕES E RELAÇÕES NAS ONTOLOGIAS BIOMÉDICAS

O grande obstáculo para um maior formalismo das ontologias da área biomédica são os problemas relacionados ao campo definição e às relações dos termos. O campo definição da maioria dessas ontologias, quando preenchido, é escrito em linguagem natural, sendo de difícil "entendimento" por máquinas. Além disso, as relações que deveriam construir a semântica dos termos são na maioria das vezes escassas, geralmente três ou quatro, havendo um sobrecarga das relações existentes. Essas relações são geralmente organizadas de forma hierárquica, somente descrevendo relações de: sinonímia (mesmo significado), hiperonímia (significado mais amplo) e hiponímia (significado mais restrito) (FREITAS; SCHULZ; MORAES, 2009).

Contudo o campo definição das ontologias biomédicas, embora, algumas vezes, escrito em linguagem natural, descreve características importantes que poderiam ser utilizadas para enriquecê-las, como possíveis relações e conceitos implícitos. Na grande maioria dessas ontologias esse campo é utilizado para complementar o conhecimento que não foi explicitado através dos conceitos e relações. Isso ocorre principalmente devido a pouca quantidade de relações e conceitos para representar o domínio.

Utilizando os conceitos expressos por Dahlberg (1978; 1981) sobre a definição é possível na *Gene Ontology*, por exemplo, encontrar diversas definições com informações implícitas (como relações e termos) que poderiam ser utilizadas para enriquecê-la, como ilustrado nos exemplos a seguir.

Termo: *Mitochondrion* (GO: 0005739)

Definição: *"A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration."*

A partir dessa definição, é possível extrair três relações: ***is*** *semiautonomous, self replicating organelle;* ***occurs in*** *varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells;* ***the site of*** *tissue respiration.*

Termo: *Endoplasmic Reticulum* (GO: 0005783)

Definição: *"The irregular network of unit membranes, visible only by electron microscopy, that occurs in the cytoplasm of many eukaryotic cells."*

Nessa definição, é possível extrair duas relações: ***is a*** *network of unit membranes, visible only by electron microscopy;* ***occurs in*** *cytoplasm of many eukaryotic cells.*

Essas informações implícitas estão sendo objeto de pesquisa em muitos grupos, principalmente os ligados ao consórcio OBO. Esses grupos estão realizando diversas análises com objetivo resgatar esse conhecimento, trabalhá-lo e validá-lo. Esse conhecimento, extraído a partir das definições, pode vir a complementar ou servir para corrigir inconsistências nas ontologias (MUNGALL, 2004; MUNGALL *et al.*, 2010; GENE ONTOLOGY, 2010d).

Além dessas iniciativas, percebe-se também um grande esforço do projeto da *Gene Ontology* e de todo o consórcio OBO de padronização das definições de suas ontologias a fim de que as mesmas possam ser trabalhadas por máquinas, permitindo um maior poder de inferência. Um exemplo é o padrão para definição de termos relacionados aos processos metabólicos que são definidos da seguinte forma na GO (GENE ONTOLOGY, 2010b):

Termo: [x] *metabolic process*

Definição Padrão: *The chemical reactions and pathways involving* [x], [descrição de x].

Outros exemplos de padrões de definição definidos pela GO são descritos no Anexo A desse trabalho. Embora esses padrões tenham sido criados inicialmente para a ontologia GO, no decorrer desse trabalho verificou-se que eles também estão sendo utilizados em outras ontologias do consórcio OBO.

No que tange às relações, a grande maioria das ontologias biomédicas possui poucos tipos de relações. A GO, maior ontologia do consórcio OBO, por exemplo, possui somente quatro tipos oficiais de relações para ligação (conexão) entre seus termos (GENE ONTOLOGY, 2010e): *is_a, part_of, has_part e regulates*.

Relação *is_a*

A existência da relação *is_a* entre dois termos, A *is_a* B, significa dizer que A é um subtipo de B. Essa relação é transitiva, portanto ao dizer que existe duas relação: A *is_a* B e B *is_a* C, pode-ser inferir uma terceira, A *is_a* C (GENE ONTOLOGY, 2010e). As relações da GO, e de outras ontologias biomédicas, sempre devem ser utilizadas para relacionar classes de entidades ou fenômenos como, por exemplo, homem *is_a* mamífero, não podendo ser aplicadas a exemplos específicos de uma dada classe (instanciação), como por exemplo: Antônio *is_a* homem (GENE ONTOLOGY, 2010e). A Figura 5 ilustra um exemplo da relação *is_a* na *Gene Ontology*.

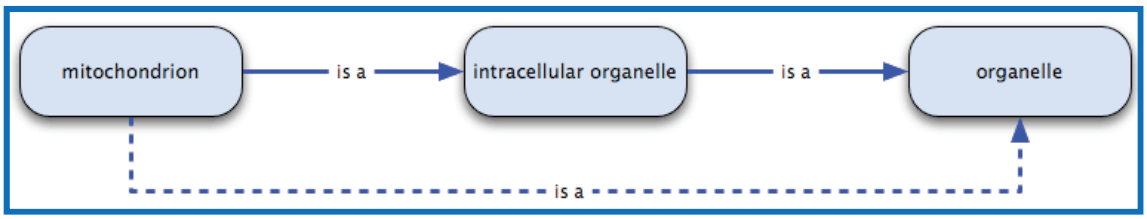

Figura 5: Exemplo da relação *is_a* na GO
Fonte: Extraído da Gene Ontology (2010e)

Relação *part_of*

A existência da relação *part_of* entre dois termos, A *part_of* B, significa dizer que todo A é necessariamente parte de B. Isso quer dizer que se existe A necessariamente implica na existência de B, porém a existência de B não implica necessariamente na existência de A (GENE ONTOLOGY, 2010e). Assim, como a relação *is_a*, essa relação é transitiva, portanto ao dizer que existe duas relações: A *part_of* B e B *part_of* C, pode-se inferir uma terceira, A *part_of* C (GENE ONTOLOGY, 2010e). A Figura 6 ilustra um exemplo da relação *part_of* na *Gene Ontology*.

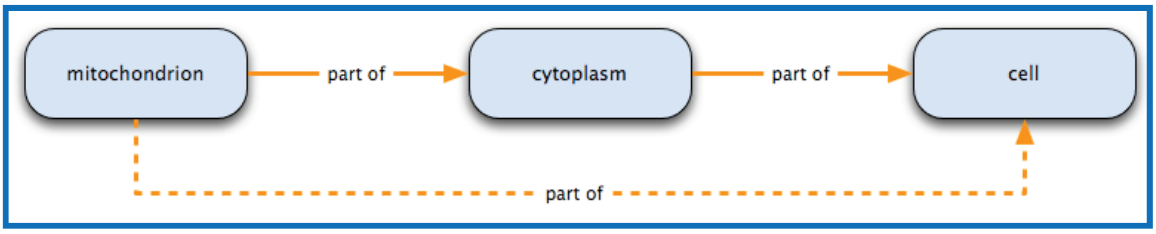

Figura 6: Exemplo da relação *part_of* na GO
Fonte: Extraído da Gene Ontology (2010e)

Relação *has_part*

A relação *has_part* é o complemento lógico da relação *part_of*. Entretanto, esse tipo de relação só existe na *full* GO. A relação *has_part* entre dois termos, A *has_part* B, significa dizer que todo A tem necessariamente como parte B, ou seja, caso A exista implica necessariamente na existência de B. Porém, dada a existência de B não se pode afirmar que A existe (GENE ONTOLOGY, 2010e). Essa relação também possui a propriedade de transitividade, portanto, se existem duas relações: A *has_part* B e B *has_part* C, pode-se inferir uma terceira, A *has_part* C (GENE ONTOLOGY, 2010e). A Figura 7 ilustra um exemplo da relação *has_part* na *Gene Ontology*.

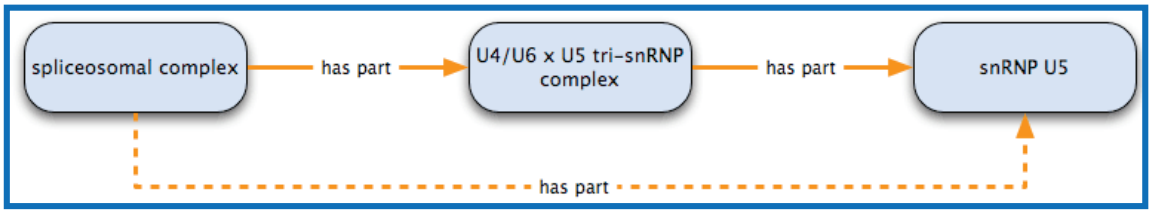

Figura 7: Exemplo da relação *has_part* na GO
Fonte: Extraído da Gene Ontology (2010e)

Relação regulates

A existência da relação *regulates* entre dois termos, A *regulates* B, significa dizer que quando A está presente necessariamente regula B, porém B pode existir sem estar necessariamente regulado por A. Na GO, a relação *regulates* é dividida em *positively regulates* e *negatively regulates* (GENE ONTOLOGY, 2010e). Essas relações ocorrem quando a manifestação de um processo (ex.: regulamentação de um caminho ou uma reação enzimática) ou de uma característica (ex.: tamanho de célula ou pH) é afetada diretamente por outro processo. Essa relação não é transitiva, portanto não se pode afirmar que: se A *regulates* B e B *regulates* C, A *regulates* C (GENE ONTOLOGY, 2010e). A Figura 8 ilustra um exemplo da relação *regulates* (*positively regulates* e *negatively regulates*) na *Gene Ontology*.

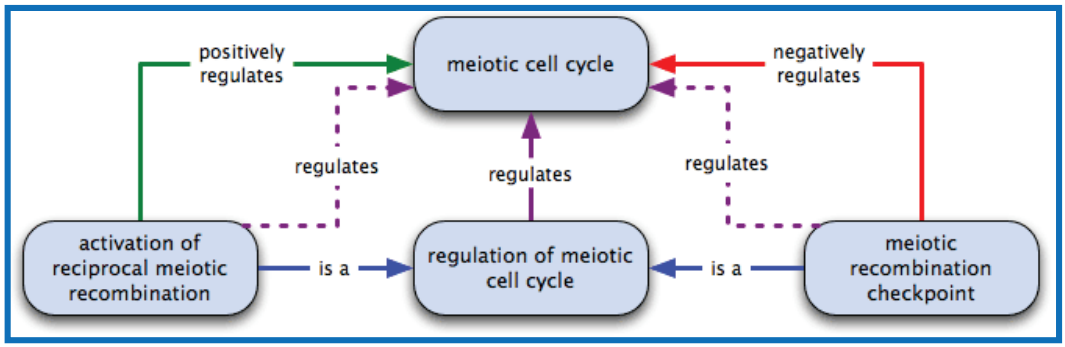

Figura 8: Exemplo da relação *regulates* na GO
Fonte: Extraído da Gene Ontology (2010e)

A pouca variedade de relações na GO e em outras ontologias do consórcio OBO pode levar a uma sobrecarga das relações existentes, o que pode ocasionar diversos problemas como, por exemplo, o uso equivocado das relações. Com o intuito de melhorar a expressividade e qualidade das ontologias biomédicas e promover uma maior padronização e documentação das relações existentes, o consórcio OBO criou o *OBO Relation Ontology* (RO) (SMITH *et al.*, 2005; OBO, 2010c). O *OBO Relation Ontology* é um padrão para definir as relações das ontologias do consórcio OBO. Esse padrão define formalmente cada tipo de relação com a finalidade de solucionar o problema de interoperabilidade e convergência entre as ontologias, conforme descrito por SMITH *et al.* (2005, p.2):

> *"O problema é que quando as ontologias da OBO e similares incorporam tais relações, isso, normalmente, ocorre por caminhos informais, muitas vezes não fornecendo nenhuma definição, fazendo com que as interconexões lógicas entre as várias relações empregadas não sejam claras, e até mesmo as relações is_a e part_of nem sempre sejam utilizadas de forma coerente, tanto dentro como entre ontologias."*.

Na sua versão inicial, a *Relation Ontology* contava com 10 relações (SMITH *et al.*, 2005; OBO, 2010c):

- Relações Fundacionais (Fundamentais):
  - *is_a*
  - *part_of*
- Relações Espaciais:
  - *located_in*
  - *contained_in*
  - *adjacent_to*

- Relações Temporais:
  - *transformation_of*
  - *derives_from*
  - *preceded_by*
- Relações de Participação:
  - *has_participant*
  - *has_agent*

Atualmente a *Relation Ontology* incorporou novas relações como: *integral_part_of* e *proper_part_of* e está em processo de análise e incorporação, pela OBO, de diversas outras (COSTA, 2009; OBO, 2010c).

Além da *Relation Ontology*, surgiram outros projetos com o objetivo de melhorar a expressividade e qualidade das ontologias biomédicas, como, por exemplo, *The Arabidopsis Information Resource* – TAIR (2010). Esse projeto é membro do *Gene Ontology Consortium* sendo o mantenedor de um banco de dados com informações sobre a genética e biologia molecular do modelo de planta *Arabidopsis Thaliana* (BERARDINI *et al.*, 2004; The Arabidopsis Information Resource, 2010). A grande vantagem dessa ontologia em relação a GO é que, enquanto o projeto da *Gene Ontology* utiliza somente quatro tipos de relações (*is_a, part_of, has_part e regulates*) para descrever o relacionamento entre seus conceitos, o projeto TAIR prevê a utilização de vinte e um tipos de relações (*has, located_in, involved_in, functions_as, expressed_in, related_to, functions_in, is_subunit_of, constituent_of, expressed_during, required_for, not_involved_in, regulates, not_expressed_in, is_down-regulated_by, expressed_only_in, not_functions_as, not_located_in, represses, expressed_only_during, not_required_for*) (BERARDINI *et al.*, 2004; SALES, 2006).

3.4 RELAÇÃO *IS_A*

3.4.1 DEFINIÇÃO DA RELAÇÃO *IS_A*

A relação é_um (*is_a*), também conhecida como relação hierárquica, genérica ou de generalização/especialização, atribui uma semântica de tipos e subtipos (pai-filho) ao

relacionar conceitos (termos ou classes) e subconceitos (subtermos ou subclasses). A relação *is_a* permite a construção de uma hierarquia de conceitos, nas terminologias, ao relacionar conceitos mais restritivos (hipônimos) com conceitos mais abrangentes (hiperônimos) (BRACHMAN, 1983; GUARINO, 1998b). As relações de *is_a* podem ser formadas através da associação de conceitos mais abrangentes (genéricos) como, por exemplo, Homem *is_a* Ser Humano ou associando um conceito abrangente (genérico) com um conceito individual, também chamado de instanciação[4], como por exemplo, Miguel *is_a* Homem (BRACHMAN, 1983). Vários aspectos devem ser considerados ao se criar uma relação de *is_a* (BRACHMAN, 1983):

- A relação de *is_a* deve sempre relacionar um conceito menos geral com um mais geral.
  - Ex.: Mulher *is_a* Ser Vivo

- A relação de *is_a* deve ser construída com base na estruturação dos conceitos e não na estruturação de frases possíveis em um domínio.
  - Ex.: Não se devem modelar as frases em domínio e sim os conceitos, pois ao se tentar modelar as frases pode-se gerar erros como explicitação de relações sem sentido (semântica) (céu *is_a* azul, por causa da frase "céu é azul").

- A relação de *is_a* deve ser sempre definida entre duas classes e essa relação será um qualificador para todos os subconceitos de um dado conceito.
  - Ex.: Não poderia, por exemplo, existir a relação Animal_que_Nada *is_a* Peixe, porque existem animais que nadam que não são peixes.

- A relação de *is_a* para ser representada deve ser realmente necessária para um domínio.
  - Ex.: Nem todas as relações de um domínio podem ser úteis como, por exemplo, explicitar cada instância de um domínio como: Miguel is_a Ser Vivo, João is_a Ser Vivo e etc. Pode torná-lo muito complexo, o correto nesse caso é representar a relação com classes genéricas como, por exemplo: Pessoa *is_a* Ser Vivo.

- A relação *is_a* sempre produz uma afirmação verdadeira para todo domínio.

[4] Relações de instanciação ou instâncias (casos particulares de classes) não são permitidas na maioria das ontologias biomédicas.

- Ex.: Ao ser inserir a relação Pessoa *is_a* Ser Vivo, isso passa a ser uma verdade para todo o domínio.

Completando esses aspectos para a criação de *is_a*, Guarino (1998b) afirma que essa relação deve ser sempre analisada antes de ser utilizada, pois a sua sobrecarga pode causar problemas à representação do domínio como, por exemplo: confusão e redução de sentido da relação, generalização excessiva de alguns conceitos, vínculos (*links*) duvidosos entre conceitos e confusão de papéis taxonômicos.

Na área de Ciência da Informação a relação *is_a* é definida por Dahlberg (1978; 1981) como uma relação hierárquica podendo ser de três tipos (gênero/espécie, espécie/espécie e espécie/indivíduos) ou como uma relação genérica que relaciona conceitos mais gerais a conceitos mais específicos (construídos através da inserção de características que especificam esses conceitos).

No âmbito biomédico, Smith *et al.* (2005) descreve a relação de *is_a* como a principal responsável pela hierarquia dos conceitos. Essa relação pode ser descrita através da associação à teoria dos conjuntos onde um subconjunto (no caso da relação *is_a*, um subconceito) está contido (pertence) a um conjunto mais geral (no caso da relação *is_a*, um conceito) (SMITH *et al.*, 2005). Smith *et al.* (2005) afirmam que a relação de *is_a* deve possuir duas características. A primeira característica é que a relação de *is_a* deve ser consistente em qualquer tempo, não permitindo a representação de relações que só ocorrem em um determinado estágio de desenvolvimento, por exemplo. A segunda característica é de que as relações devem somente estabelecer vínculos entre "verdades absolutas" (classes concretas) do contexto biológico (ex.: genes, processos e funções) (SMITH *et al.*, 2005). Ainda segundo Smith *et al.* (2005), a relação *is_a* pode ser definida formalmente de duas formas:

> **1. C *is_a* C_1 =** [definição] para todo c, t, se c *instance_of* C em t então c *instance_of* C_1 em t, onde C e C_1 são classes contínuas (*continuants*), t é um dado tempo e c é uma instância de uma classe contínua.
>
> **2. P *is_a* P_1 =** [definição] para todo p, se p *instance_of* P então p *instance_of* P_1, onde P e P_1 são classes de processos e p é uma instância de processo.

A relação *is_a* é a principal relação das terminologias, pois é através dela que é estruturada a hierarquia dos conceitos, atribuindo com isso mais semântica e expressividade. Porém, em grande parte das terminologias biomédicas, não somente ligadas ao consórcio

OBO, é possível verificar grandes falhas na utilização dessa relação como, por exemplo, relações de *is_a* sendo utilizadas de forma inconsistente e também representando outros tipos de relações como, *part_of* (KÖHLER *et al.* 2006; HOGAN, 2009).

3.4.2 PADRÕES LINGUISTICOS ASSOCIADOS À RELAÇÃO *IS_A*

Os métodos para extração de hiperonímia e hiponímia podem ser utilizados para obtenção da relação *is_a* na construção ou manutenção de domínios. Esses métodos podem ser divididos em dois tipos: *top-down* e *bottom-up* (FREITAS, 2007). Os métodos *top-down* extraem a informação a partir de padrões linguísticos pré-definidos descritos através de regras (MORIN; JACQUEMIN, 2003; FREITAS, 2007). A desvantagem desse tipo de método é a análise necessária para identificação e criação das regras, o que pode em alguns casos ser bastante complexo e requerer muito tempo (MORIN; JACQUEMIN, 2003; FREITAS, 2007). Como principal representante do método *top-down* pode-se citar o trabalho de Hearst (1992; 1998) (MORIN; JACQUEMIN, 2003; FREITAS, 2007). Os métodos *bottom-up* extraem a informação através da classificação das palavras de um corpus por meio de algoritmos de agrupamento baseado na similaridade entre os contextos em qual a palavra está inserida (MORIN; JACQUEMIN, 2003; FREITAS, 2007). A desvantagem desse tipo de método é que, na maioria das vezes, os agrupamentos não são rotulados, ou seja, não é possível verificar com precisão as relações semânticas no grupo e entre os grupos, dificultando o processo de extração (MORIN; JACQUEMIN, 2003; FREITAS, 2007).

Para o escopo desse trabalho, só nos interessa descrever com mais detalhes os métodos *top-down* uma vez que eles permitem a extração da relação de hiperonímia/hiponímia de forma mais precisa, ou seja, expressam com maior clareza as relações semânticas encontradas entre os conceitos (FREITAS, 2007). O padrão de Hearst (1992; 1998), principal representante do método *top-down*, se propõe a identificar e extrair relações de hiperonímia (relações semânticas que, em alguns casos, podem vir a ser uma relação *is_a*) através de padrões léxico-sintáticos, pré-definidos, aplicados aos conjuntos de sentenças existentes no corpus (FREITAS, 2007). Para isso, Hearst (1992; 1998) propõe seis tipos de regras:

1. NP_0 *such as* NP_1 {, NP_2..., (*or* | *and*) NP_i}, onde para cada i>0 tem-se um Hipônimo (NP_i, NP_0).

Ex.: […] *physiologic systems* ***such as*** *stress response and immune response* […]

Hipônimo (*stress response*, *physiologic systems*)

Hipônimo (*immune response*, *physiologic systems*)

2. *such* NP *as {*NP_1*,}* {(or | and)}* NP_i, onde para cada NP_i tem-se um Hipônimo (NP_i, NP).

   Ex. [...] ***such*** ontologies *as Gene Ontology, INOH and Brenda* […]

   Hipônimo (*Gene Ontology*, o*ntology*)

   Hipônimo (*INOH*, o*ntology*)

   Hipônimo (*Brenda*, o*ntology*)

3. NP_1 {, NP_i}* {,} *or other* NP, onde para cada NP_i tem-se um Hipônimo (NP_i, NP).

   Ex.: […] *INOH, Brenda, Pathway* ***or other*** *biomedical ontologies* […]

   Hipônimo (*INOH*, *biomedical ontology*)

   Hipônimo (*Brenda*, *biomedical ontology*)

   Hipônimo (*Pathway*, *biomedical ontology*)

4. NP_1 {, NP_i}* {,} *and other* NP, onde para cada NP_i tem-se um Hipônimo (NP_i, NP).

   Ex.: […] *plasma membrane*, *nucleus* ***and other*** *ontology subcellular structures* […]

   Hipônimo (*plasma membrane*, *subcellular structures*)

   Hipônimo (*nucleus*, *subcellular structures*)

5. NP {,} *including* {NP_0,}* {*or* | *and*} NP_i, onde para cada NP_i tem-se um Hipônimo (NP_i, NP).

   Ex.: [...] *Ontologies of OBO consortium,* ***including*** *Pathway, and ChEBI* […]

   Hipônimo (*Pathway*, *Ontology of OBO consortium*)

   Hipônimo (*ChEBI*, *Ontology of OBO consortium*)

6. NP {,} *especially* {NP_0,}* {*or* | *and*} NP_i, onde para cada NP_i tem-se um Hipônimo (NP_i, NP).

   Ex.: [...] *biomecial terms,* ***especially*** *DNA-repair, and biological process* […]

   Hipônimo (*DNA-repair*, *biomecial terms*)

   Hipônimo (*Biological Process*, *biomecial terms*)

As regras 1, 2 e 3, descritas anteriormente, foram levantadas manualmente por Hearst através da observação e análise de corpora (HEARST, 1998; FREITAS, 2007). As demais

regras foram construídas através de um procedimento descrito por Hearst divido em 4 etapas (HEARST, 1998; FREITAS, 2007):

1. Decidir a relação lexical “R” de interesse.
2. Identificar uma lista de pares de palavras, do WordNet (MILLER, 1995), que se relacionam utilizando a relação “R”.
3. Extrair de um corpus sentenças em que esses pares de palavras aparecem, analisando e registrando o contexto lexical e sintático.
4. Encontrar os pontos em comum entre os contextos analisados, criando regras a fim de que se possa generalizá-las e transformá-los em um padrão.

O procedimento proposto por Hearst (1998), descrito anteriormente, pode ser adaptado e expandido para a descoberta de diferentes regras que podem ser utilizadas para a criação de padrões linguísticos (FREITAS, 2007). Além disso, através de procedimentos parecidos com esse, é possível a descoberta de padrões em diversos idiomas como, por exemplo, Português (FREITAS; QUENTAL, 2007) e Francês (MORIN; JACQUEMIN, 2003) (FREITAS, 2007).

3.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou diversos conceitos importantes relacionados ao campo definição e as relações das terminologias de uma forma geral, sendo especializado no contexto das ontologias biomédicas e na relação *is_a*. Essa discussão foi proposta com a finalidade de se evidenciar a importância desse campo e desse tipo de relação para essas terminologias, principalmente as ontologias.

Além disso, esse capítulo também demonstrou, a partir dos princípios das definições estudados, que o campo definição das ontologias biomédicas possui diversas informações implícitas, entre elas, relações de *is_a* e novos termos que poderiam vir a ser úteis para aumentar a semântica das ontologias favorecendo os processos de reúso e anotação.

O próximo capítulo irá discutir os aspectos relacionados à extração de informação, mostrando características, abordagens, enfoques e ferramentas, tendo como foco principal a

extração de informação no contexto biomédico, que será tratada com mais detalhes em uma de suas subseções.

CAPÍTULO 4: EXTRAÇÃO DE INFORMAÇÃO

Este capítulo apresenta aspectos relacionados à extração de informação, como características, abordagens, enfoques e ferramentas com foco principalmente no domínio biomédico. Além disso, ao final desse capítulo são apresentadas algumas iniciativas de extração de informação na área biomédica, que serviram como apoio para o desenvolvimento da abordagem proposta nessa dissertação.

4.1 CONCEITOS RELACIONADOS À EXTRAÇÃO DE INFORMAÇÃO

O processo de extração de informação (EI) é constituído de 2 passos (ANANIADOU; McNAUGHT, 2006):

1. Aplicar processamento de linguagem natural a um corpus (conjunto de documentos) e extrair dele fatos essenciais e tipos pré-definidos.
2. Representar cada fato como um modelo (*template*) de saída, cujos campos (*slots*) do modelo são preenchidos com base nas informações encontradas.

Objetivamente, a extração de informação pode ser definida como uma aplicação de processamento de linguagem natural (PLN) que tem como objetivo a extração de informações (conhecimento) do corpus (ANANIADOU; McNAUGHT, 2006). Ela se diferencia do conceito de recuperação de informação que trabalha com o auxílio de índices em um conjunto de documentos e tem por objetivo trazer uma resposta a uma busca (ANANIADOU; McNAUGHT, 2006). Por esse motivo, os mecanismos de recuperação de informação são conhecidos como "motores de busca". Exemplos de mecanismos de recuperação de informação são Google[5], Bing[6] e Yahoo[7] (ANANIADOU; McNAUGHT, 2006). A Tabela 4 extraída do livro *Text Mining for Biology and Biomedicine* (ANANIADOU; McNAUGHT, 2006, p.145) resume a diferença entre os dois conceitos.

A extração da informação, normalmente, ocorre em duas etapas. Na primeira etapa ocorre o processamento da linguagem natural através de mecanismos de processamento de linguagem. Na segunda etapa são aplicados abordagens, enfoques e técnicas com a finalidade de extrair informações. Embora essas etapas possuam suas particularidades e finalidades, em muitos trabalhos, elas aparecem mescladas em uma única etapa.

[5] www.google.com.br
[6] www.bing.com
[7] br.yahoo.com

Tabela 4: Diferenças entre extração de informação e recuperação de informação

Recuperação de Informação (RI)	Extração de Informação (EI)
Retorna documentos	Retorna fatos
Classificação de documentos com a finalidade de separar os documentos relevantes dos não relevantes para responder a uma consulta.	Aplicação de processamento de linguagem natural (PLN), utilizando-se de análise do texto e síntese das estruturas de representação
Pode ser feito sem referência à sintaxe (trata a consulta e os documentos como um "pacote de palavras")	Baseado na análise sintática e semântica

Fonte: Ananiadou e McNaught (2006, p.145)

O processamento de linguagem natural é o conjunto de abordagens, algoritmos, mecanismos e técnicas para analisar (processar) e anotar os corpora escritos em linguagem natural. Segundo CIMIANO (2006) a linha de produção (*pipeline*) para processamento de linguagem natural pode ser divida em quatro etapas ilustradas na Figura 9 e detalhadas a seguir.

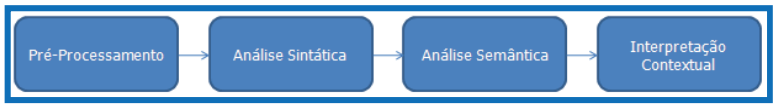

Figura 9: Linha de produção para processamento de linguagem natural
Fonte: Cimiano (2006, p.37)

- **Pré-Processamento:** essa etapa tem por objetivo preparar o corpus e anotá-lo com informações que irão servir de base para as demais etapas. Ainda segundo Cimiano (2006) a etapa de Pré-processamento pode ser divida em cinco subetapas, conforme mostrado na Figura 10.

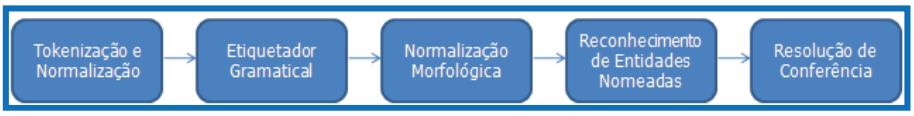

Figura 10: Subetapas do pré-processamento do corpus
Fonte: Cimiano (2006, p.37)

  - **Tokenização e Normalização:** essa subetapa tem como objetivo realizar o processo de tokenização e normalização. O processo de tokenização tem como função a separar as palavras ou sentenças do corpus em tokens. O processo de normalização tem como função padronizar os formatos do corpus, transformando, por exemplo, datas e unidades monetárias em um formato único.

- **Etiquetador Gramatical:** essa subetapa tem por objetivo anotar cada token de palavras com sua classe gramatical como, por exemplo, advérbio, adjetivo, substantivo, pronome e etc.
- **Normalização Morfológica:** essa subetapa tem por objetivo reduzir as palavras do corpus a uma forma única. O processo de normalização morfológica pode ser feito através de radicalização (*stemming*) ou lematização.
- **Reconhecimento de Entidades Nomeadas:** essa subetapa tem por objetivo o reconhecimento de nomes (entidades) no corpus, atribuindo, aos nomes identificados, classes como, objetos, pessoas, carros, empresas, endereços, marcas e etc. Geralmente nessa etapa adota-se um dicionário com os nomes e as classes mais comuns para auxiliar no processo.
- **Resolução de Conferência:** essa subetapa tem por objetivo associar nomes que representam uma mesma entidade, como, por exemplo: GRECO e Grupo de Engenharia do Conhecimento.

- **Análise Sintática:** essa etapa tem por objetivo a análise sintática dos componentes que constituem a sentença (ex.: sujeito, verbo e objeto direto e indireto). A Figura 11 ilustra uma árvore de *parser* (CIMIANO, 2006), um exemplo de estrutura utilizada para auxiliar na análise sintática das sentenças do corpus.

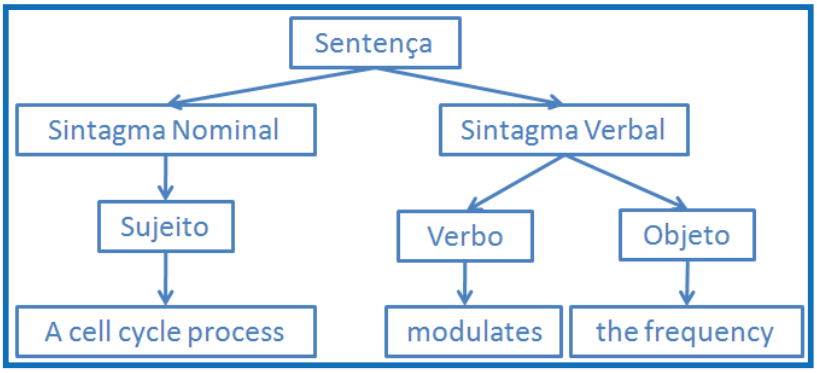

Figura 11: Exemplo de análise sintática utilizando a estrutura de árvore de *parser*

Fonte: Adaptado de Cimiano(2006, p.41)

- **Análise Semântica:** essa etapa tem por objetivo fazer uma análise e validação semântica das sentenças. Para isso, essa etapa pode contar com o auxílio de ontologias ou dicionários especializados como, por exemplo, o WordNet (MILLER, 1995). Muitas abordagens utilizam essa etapa para extração de

informação propriamente dita. Isso é possível, pois a análise semântica valida à estrutura das sentenças e com isso permitir a identificação dos componentes.

- **Interpretação Contextual:** essa etapa tem por objetivo a retirada de ambiguidades do corpus através da análise do contexto em que as sentenças estão inseridas. Ex.: "manga" que pode ser a fruta ou a manga da camisa.

Após a análise e anotação do corpus através do processamento de linguagem natural, é aplicada a segunda etapa do processo de extração de informação que tem a função de identificar as informações desejadas no corpus. Para isso são utilizados diversos enfoques. Os principais enfoques utilizados para extração de informações no contexto biomédico, segundo Ananiadou e McNaught (2006), são:

- **Enfoque baseado em dicionário** utiliza recursos existentes na terminologia para extrair informação do corpus. Esse enfoque utiliza-se de um dicionário de palavras para fazer a marcação dos conceitos no corpus, fazendo uma correspondência com as palavras do dicionário. As principais vantagens desse enfoque é a sua facilidade de implementação (basta ter um dicionário de palavras) e também a possibilidade de correspondência perfeita (todas as palavras do dicionário são marcadas no corpus). Porém esse enfoque não consegue identificar neologismos e sua precisão diminui em corpus com muitas palavras ambíguas.

- **Enfoque baseado em regras** utiliza diversos padrões de formação para extrair informação do corpus. Nesse enfoque são definidos padrões de formação que podem utilizar desde simples regras de identificação de substantivos até complexas análises ortográficas para encontrar os conceitos relevantes no corpus. A principal vantagem desse enfoque, em relação aos demais, é que ele possui uma melhor precisão. Entretanto, esse enfoque requer que sejam construídas regras para a sua utilização o que necessita de conhecimento prévio do corpus, com o agravante de que as regras criadas possam ser restritas a um domínio específico.

- **Enfoque baseado em aprendizagem de máquina** utiliza técnicas de aprendizagem de máquina para extração de informação do corpus. A principal vantagem desse enfoque é a possibilidade de inferência do conhecimento, a partir do treinamento recebido. O principal desafio é encontrar um bom conjunto amostral para realização do treinamento.

- **Enfoque baseado em estatística** utiliza técnicas estatísticas que se baseiam em diferentes distribuições das colocações do texto no corpus para extração de informação. A principal vantagem desse enfoque é que ele é mais facilmente adaptado a outros domínios que os demais e a principal desvantagem é a dificuldade em definir fórmulas estatísticas adequadas para sua aplicação. Além disso, esse enfoque possui seu desempenho limitado quando aplicado a corpus restrito.

A Tabela 5, construída com base nos conceitos expressos no livro de Ananiadou e McNaught (2006), faz uma comparação entre os enfoques mostrando as vantagens e desvantagens.

Tabela 5: Comparação entre os enfoques de extração de informação

Enfoque	Vantagem	Desvantagem
Dicionário	Possui correspondência perfeita com as palavras do dicionário	Limitado às palavras do dicionário, não identifica novos neologismos Termos ambíguos diminuem a precisão
Regras	Melhor precisão do que os outros enfoques	Criação de regras pode ser limitada a um domínio específico
Aprendizado de Máquina	Inferência de conhecimento no corpus	Dependente do treinamento recebido
Estatísticas	Mais facilmente adaptado para diferentes domínios do que os demais enfoques.	Definir fórmulas estatísticas adequadas para sua aplicação no corpus Limitada quando o corpus é restrito

Fonte: Baseado em Ananiadou e McNaught (2006)

Ananiadou e McNaught (2006) recomendam que o processo de reconhecimento de um termo seja executado por etapas para que seja possível extrair relações e propriedades realmente relevantes para um determinado contexto. Nessa linha, afirmam que, para extração de um termo (representante de um conceito em um domínio), é preciso passar por três etapas (conforme Figura 12). A primeira etapa é o reconhecimento de uma ou mais palavras que representam um conceito. A segunda etapa é a classificação do termo (conceito) em classes do domínio (como, por exemplo, tecidos, proteínas, enzimas e genes). A terceira e última etapa é o mapeamento do termo no domínio, ou seja, o estabelecimento de relações do termo com os outros termos (conceitos) do domínio. Essa divisão em etapas permite uma melhor eficiência na extração de informação. Entretanto, a maioria das abordagens do domínio biomédico não

utiliza uma abordagem sistemática por etapas, e mescla todas as etapas em uma única, comprometendo a qualidade do processo (ANANIADOU; McNAUGHT, 2006).

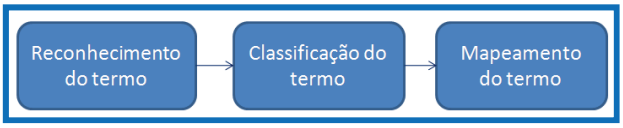

Figura 12: Etapas para extração de Termos (Conceitos)
Fonte: Ananiadou e McNaught (2006, p.73)

4.2 FERRAMENTAS PARA ANÁLISE LINGUÍSTICA E EXTRAÇÃO DE INFORMAÇÃO

No decorrer desse trabalho foram testadas diversas ferramentas e mecanismos para a análise linguística e para extração de informação. A seguir, cinco desses mecanismos são descritos e a escolha pelo GATE (2010) e NLTK (2010) neste trabalho é justificada.

4.2.1 GATE

O framework *General Architecture for Text Engineering* (GATE) (2010) foi lançado originalmente em 1996. Posteriormente, em 2002, foi relançado tendo sido totalmente redesenhado e reescrito, atualmente esse framework é utilizado por milhares de pessoas (BONTCHEVA *et al.*, 2004). O framework é composto por diversos recursos, tendo sua arquitetura projetada para permitir a utilização ou desenvolvimento de diferentes componentes da engenharia de linguagens (CUNNINGHAM *et al.*, 2002). As principais vantagens do framework GATE são (CUNNINGHAM *et al.*, 2002; BONTCHEVA *et al.*, 2004):

- Desenvolvimento baseado em componentes.
- Aplicações independentes dos recursos, possibilitando a fácil inserção ou remoção de recursos.
- Possui código-fonte aberto e documentado.
- Separação entre as tarefas de baixo nível (como, por exemplo, visualização de dados) e de alto nível (como, por exemplo, tarefas de processamento de

linguagem).

- Compatibilidade com diferentes padrões e formatos.

Para trabalhar o corpus o framework GATE é composto por três tipos de recursos (CUNNINGHAM *et al.*, 2002; BONTCHEVA *et al.* 2003):

- Recursos de Linguagem: são os recursos linguísticos utilizados pelo GATE. Ex.: corpus e ontologias.

- Recursos de Processamento: são os recursos utilizados para processamento das informações. Ex.: Parser.

- Recursos de Visualização: são os recursos utilizados para visualização das informações. Ex.: recursos associadas ao GUI.

Esses recursos disponibilizados no GATE são conhecidos como CREOLE (*Collection of REusable Objects for Language Engineering*). A utilização desses recursos permite, ao usuário, construir complexas linhas de produção (*pipelines*) formando diversas arquiteturas de engenharia de linguagem, possibilitando inclusive a utilização de recursos externos como, por exemplo, o WordNet (MILLER, 1995) (CUNNINGHAM *et al.*, 2002).

Para extração de informação o principal recurso do framework GATE é o ANNIE (*A Nearly New Information Extraction System*). O recurso ANNIE consiste basicamente nos seguintes componentes (BONTCHEVA *et al.* 2003; CUNNINGHAM *et al.*, 2010):

- **Tokeniser:** esse componente divide o corpus em tokens (como palavras, espaços e pontuações) através de expressões regulares.

- **Sentence Splitter:** esse componente divide o corpus em sentenças utilizando como apoio listas definidas no dicionário do *Gazetteer*.

- **Etiquetador Gramatical (Part of Speech Tagger):** esse componente utiliza as sentenças divididas no *Sentence Splitter* para fazer a marcação do corpus em classes gramaticais (ex.: adjetivos, advérbios substantivos, pronomes e verbos) com base em um treinamento com o corpus do *Wall Street Journal.*

- **Gazetteer:** esse componente faz a marcação das palavras do corpus a partir de um dicionário próprio que contém listas de nomes (conceitos) associados às categorias.

- **Transducer:** esse componente utiliza-se de regras pré-definidas pelo GATE ou criadas pelo usuário, para marcação do corpus. Nele é possível criar complexos programas para reconhecimento de categorias, através da linguagem JAPE (*Java Annotation Patterns Engine*).

- **Orthomatcher:** esse componente verifica e completa as marcações dos demais componentes utilizando listas com co-referência entre nomes (conceitos).

4.2.2 NLTK

O NLTK (2010) (*Natural Language Toolkit*) foi criado em 2001 pelo Departamento da Computação e Ciência Informação da Universidade da Pensilvânia. NLTK é um conjunto de recursos, para processamento de linguagem natural, executados a partir de programas desenvolvidos em Python[8] (BIRD; KLEIN; LOPER, 2009).

NLTK possui quatro objetivos principais (BIRD; KLEIN; LOPER, 2009):

- **Simplicidade:** o NLTK possui uma estrutura simples e intuitiva, permitindo até os usuários pouco experientes a construção de blocos de funcionalidades e conhecimento prático em PLN (Processamento de Linguagem Natural).

- **Consistência:** o NTLK possui métodos com nomes que visam facilitar a compreensão e o entendimento sobre suas funcionalidades. E também estruturas de dados com padrões consistentes.

- **Extensibilidade:** o NLTK possui uma estrutura que visa facilitar a inclusão, remoção ou implementação de componentes permitindo a construção de aplicações de acordo com a necessidade do usuário.

- **Modularidade:** o NLTK possui independência de funcionamento entre os seus componentes.

[8] www.python.org

O NLTK disponibiliza diversos algoritmos para processamento do corpus e fornece diversas bibliotecas para tratamento do corpus em diferentes idiomas. Além disso, permite que seja integrado com diversas outras ferramentas como, por exemplo, WordNet (MILLER, 1995) (BIRD; KLEIN; LOPER, 2009). A Tabela 6, extraída de Bird, Klein e Loper (2009), mostra uma lista dos principais módulos do NLTK.

Tabela 6: Principais Módulos do NLTK

Tarefa	Módulo NLTK	Exemplo de Funcionalidades
Acessar os corpora	nltk.corpus	interfaces padronizadas para corpora e léxicos
Processamento de *String*	nltk.tokenize, nltk.stem	tokenizers, tokenizers de sentença, normalizações morfológicas (ex.: stemmer)
Testes Estatísticos	nltk.collocations	teste t, qui-quadrado
Etiquetador grammatical (*Part-of-speech tagger*)	nltk.tag	n-gram, backoff, Brill, HMM, TnT
Classificação	nltk.classify, nltk.cluster	árvore de decisão, naive Bayes, k-médias
Chunking	nltk.chunk	expressão regular, n-gram
Análise Sintática (*Parsing*)	nltk.parse	chart, probabilística, dependência
Interpretação semântica	nltk.sem, nltk.inference	cálculo do lambda, lógica de primeira ordem
Métricas de avaliação	nltk.metrics	precisão, revocação, correlação de coeficientes
Probabilidade e estimativa	nltk.probability	distribuições de frequência, distribuições de probabilidades
Aplicações	nltk.app, nltk.chat	parsers, WordNet, chatbots
Trabalho no campo linguístico	nltk.toolbox	manipulação de dados

Fonte: Bird, Klein e Loper (2009)

4.2.3 ONTO LT

OntoLT é um *plugin* desenvolvido para ser utilizado em conjunto com o framework Protégé (STANFORD, 2010). O objetivo desse *plugin* é a construção de ontologias a partir das informações extraídas de um corpus através de técnicas de análise linguística. Para isso, o OntoLT possui regras para mapeamento das frases do corpus em conceitos (classes) e atributos (*slots*) da ontologia construída (BUITELAAR; OLEJNIK; SINTEK, 2004).

Antes de realizar a extração dos conceitos e atributos, pelo plugin OntoLT, é necessário que o corpus esteja pré-processado com informações linguísticas como, por exemplo, etiquetador gramatical e análise da estrutura das palavras. Na versão 1.0 do OntoLT esse pré-processamento é feito utilizando a ferramenta SCHUG (DECLERCK, 2002) (BUITELAAR; OLEJNIK; SINTEK, 2004; BUITELAAR; SINTEK, 2004; SINTEK; BUITELAAR; OLEJNIK, 2004). Essa ferramenta cria um XML contendo as informações anotadas no corpus através da utilização de enfoques baseados em regras que permitem a realização do pré-processamento e marcação do corpus nos idiomas inglês ou alemão (BUITELAAR; OLEJNIK; SINTEK, 2004; BUITELAAR; SINTEK, 2004; SINTEK; BUITELAAR; OLEJNIK, 2004).

Posteriormente a esse pré-processamento, são aplicadas no corpus as regras pré-definidas pelo *plugin* ou criadas pelo usuário com a finalidade de extração dos conceitos e atributos que irão formar a ontologia (BUITELAAR; OLEJNIK; SINTEK, 2004; BUITELAAR; SINTEK, 2004). Exemplos de regras pré-definidas pelo OntoLT são (BUITELAAR; OLEJNIK; SINTEK, 2004; SINTEK; BUITELAAR; OLEJNIK, 2004; SINTEK; BUITELAAR; OLEJNIK, 2004):

- **HeadNounToClass_ModToSubClass:** essa regra mapeia um substantivo, cabeça (chave), para uma classe e a combinação de seus modificadores para uma ou mais subclasses.

- **SubjToClass_PredToSlot_DObjToRange:** essa regra mapeia o sujeito para uma classe, o seu predicado é mapeado para um atributo (*slot*) e o objeto é mapeado para o intervalo desse atributo (*slot*).

Ao final do processo de extração o OntoLT gera uma ontologia que pode ser trabalhada no Protégé.

4.2.4 SAS TEXT MINER

SAS Text Miner (SAS Institute Inc, 2010) é um pacote do *SAS ® Enterprise Miner*. Esse pacote tem como principal função a extração informações e reconhecimento de padrões e tendências. Com seus mecanismos o *SAS Text Miner* pode, por exemplo, auxiliar as

organizações nas tomadas de decisões estratégicas e no reconhecimento de novas oportunidades através das tendências e padrões identificados (SAS Institute Inc, 2010).

Para usuários comuns, o *SAS Text Miner* pode ser um importante instrumento para análise e extração do corpus, permitindo entre outras funcionalidades, o auxílio para a construção de domínios (para ontologistas), a análise da correlação entre as informações extraídas, o agrupamento de um conjunto de documentos por temas ou facetas e a visualização das informações extraídas ou do corpus sobre diferentes perspectivas e prismas (SAS Institute Inc, 2010).

O *SAS Text Miner* possui diversas funcionalidades para trabalhar o corpus, além das ferramentas de pré-processamento que permite a modelagem e marcação linguística do corpus, esse pacote conta, ainda, com um amplo conjunto de funcionalidades como, por exemplo, ferramentas de análise ortográfica, ferramentas de análise de tendências, algoritmos para transformação de formatos, agrupamento de documentos, ferramentas de análise das correlações entre as informações extraídas, técnicas de aprendizagem de máquina, técnicas de extração baseada em regras e etc. (SAS Institute Inc, 2010).

Um dos grandes diferencias desse mecanismo são as diferentes formas de visualização das informações. Nesse mecanismo é permito desde a visualização de alto nível, mostrando a disposição das palavras no corpus até a visualização em contexto de baixo nível como a apresentação de uma sentença específica. Além disso, as ferramentas de visualização permitem, através de uma forma gráfica, uma avaliação completa do contexto no qual um conceito está inserido (SAS Institute Inc, 2010).

As principais vantagens do SAS Text Miner são (SAS Institute Inc, 2010):

- Suporte a diferentes formas de visualização.
- Agrupamento por categorias de conceitos.
- Utilização de processos automatizados.
- Suporte a múltiplos idiomas.
- Suporte a diferentes formatos de arquivos.

- Refinamento das informações extraídas, permitindo a identificação de novas informações não encontrada em uma primeira execução.
- Diferentes níveis de detalhamento das informações do corpus.
- Utilização de diversos enfoques para extração de informação como, por exemplo, regras para mapeamento(enfoque baseado em regras) e técnicas de aprendizagem de máquina (enfoque baseado em aprendizagem de máquina).
- Interfaces intuitivas e de fácil utilização.
- Análise e verificação ortográfica e sintática.
- Técnicas automatizadas para pré-processamento do corpus.
- Possibilidade de reconhecimento de novos conceitos no corpus.
- Reconhecimento automático de novas tendências e oportunidades.

4.2.5 TROPES

A ferramenta Tropes (CYBERLEX, 2010; CYBERLEX, 2006) tem por objetivo a mineração e extração de informação de um determinado corpus. Essa ferramenta, desenvolvida pela Cyberlex, permite fazer análises automatizadas ou semiautomatizadas em conjuntos de corpus escritos em linguagem natural utilizando-se de perspectivas semânticas e pragmáticas. Com isso, é possível garantir a qualidade e a confiabilidade das informações extraídas (CYBERLEX, 2010; CYBERLEX, 2006).

Para análise e extração do corpus o Tropes possui diversos mecanismos acoplados como (CYBERLEX, 2010; CYBERLEX, 2006):

- Mecanismos de Pré-Processamento, incluindo etiquetagem gramatical (*part-of-speech tagger*).
- Mecanismos para visualização do corpus sobre diferentes prismas como, por exemplo, filtro no corpus por categoria ou por tema.

- Mecanismos para análises automáticas do corpus, como por exemplo, mecanismos para análises das séries cronológicas no corpus.

As principais vantagens dessa ferramenta são (CYBERLEX, 2010; CYBERLEX, 2006):

- Compatibilidade com diversos idiomas, como: inglês e português.
- Algoritmos para análise linguística capazes de tratar corpus com milhares de documentos sem requerer grande quantidade de recurso de processamento da máquina.
- Utilização de algoritmos de inteligência artificial para classificação das palavras do corpus e para resolução de ambiguidades.
- Utilização de um dicionário, com mais de 500000 classificações semânticas, para marcação do corpus.
- Suporte a diferentes formatos de arquivos, como, por exemplo, pdf, doc e txt.

Para mineração e extração de informação, a ferramenta Tropes utiliza-se de diferentes análises, tais como (CYBERLEX, 2010):

- **Análise morfossintática:** utilizada para identificar a categoria morfológica (adjetivo, advérbio, substantivo, verbo, etc) das palavras do corpus.
- **Análise léxico-semântica:** utilizada para classificar o corpus em termos semânticos (palavras relevantes para o corpus).
- **Análise cognitivo-discursiva** utilizada para permitir uma análise sobre a cronologia do discurso. Essa análise permite dividir o corpus em "episódios" (categorias) promovendo um melhor entendimento da sequências das informações.

A Tabela 7 faz uma comparação entre os mecanismos de análise linguística e de extração de informação pesquisados. Nessa dissertação foram escolhidos o framework GATE e o mecanismo NLTK para tratamento da definição. O principal motivo dessa escolha foi que ambos possuem ampla variedade de ferramentas para processamento de informação, licença livre, código aberto, vasta documentação e módulos facilmente

adaptáveis para diferentes aplicações. Além disso, optou-se por utilizar os dois para aproveitar as vantagens individuas de ambos, permitindo com isso um pré-processamento mais completo.

Tabela 7: Comparação entre os mecanismos de Extração de Informação analisados

Mecanismo / Categoria	GATE	NLTK	OntoLT	SAS Text Miner	Tropes
Licença	Livre	Livre	Livre	Proprietária	Proprietária
Open Source	Sim	Sim	Não	Não	Não
Uso Amigável[9]	Sim	Sim	Sim	Sim	Sim
Interface Gráfica	Sim	Não	Sim	Sim	Sim
Documentação	Documentação no site do desenvolvedor e na internet	Documentação no site do desenvolvedor e na internet	Pouca documentação no site do desenvolvedor	Documentação no site do desenvolvedor e no manual do usuário	Documentação no site do desenvolvedor e no manual do usuário
Objetivo Principal	Analise linguística, extração e mineração	Analise linguística, extração e mineração	Analise linguística, extração de informação para criação de ontologias	Analise linguística, extração e mineração	Analise linguística, extração e mineração
Site	http://gate.ac.uk	http://www.nltk.org	http://olp.dfki.de/OntoLT/OntoLT.htm	http://www.sas.com	http://www.cyberlex.pt/produtos.html
Pré-Processamento	Sim	Sim	Não	Sim	Sim
Fácil adaptação para diferentes contextos (mudança das funcionalidades)	Sim	Sim	Não	Sim	Sim

[9]Funcionalidades intuitivas.

4.3 EXTRAÇÃO DE INFORMAÇÃO NO CONTEXTO BIOMÉDICO

A identificação de relações e interações entre genes, funções moleculares e biológicas, reações químicas, proteínas, enzimas, drogas, doenças, e outros conceitos biomédicos são importantes para pesquisas da área biomédica (PALAKAL; MUKHOPADHYAY; STEPHENS, 2005). Nesse contexto, cada vez mais a utilização de mineração e extração de informação se tornam processos necessários para agregação de novos conhecimentos às pesquisas (PALAKAL; MUKHOPADHYAY; STEPHENS, 2005). Diversas pesquisas ligadas a essa área utilizam-se de abordagens, mecanismos e ferramentas para mineração e extração de informação com intuito de extrair informação para um melhor entendimento sobre a interação entre os conceitos desse domínio (PALAKAL; MUKHOPADHYAY; STEPHENS, 2005).

Várias abordagens, mecanismos e enfoques, (AGARWAL; SEARLS, 2008; FUNDEL; KUFFNER; ZIMMER, 2007; HAKENBERG *et al.*, 2010; KRALLINGER; VALENCIA; HIRSCHMAN, 2008; RINALDI *et al.*, 2004; RINALDI *et al.*, 2007) desenvolvidos para esta área, utilizam como corpus a própria literatura biomédica como, por exemplo, teses, dissertações e artigos científicos disponibilizados na internet, principalmente os de bases especializadas como o PubMed (2010). A maioria desses trabalhos utiliza uma sequência de atividades, para extração de informação, similar à proposta por Mathiak e Eckstein (2004). Essa abordagem utiliza-se de pré-processamento para preparar o corpus para a extração de informação propriamente dita, sendo constituída de cinco etapas (Figura 13), detalhadas a seguir (MATHIAK; ECKSTEIN, 2004):

Figura 13: Abordagem para extração de informação no domínio biomédico proposta por Mathiak e Eckstein (2004)
Fonte: Mathiak e Eckstein (2004)

- **Seleção do Corpus:** essa etapa consiste em selecionar textos para a formação do corpus. Esses textos podem ser selecionados, por exemplo, a partir de sites especializados em artigos biomédicos como o PubMed ou de ferramentas de busca (ex.: Google, Yahoo e Bing).

- **Pré-Processamento:** essa etapa consiste na aplicação de técnicas e ferramentas de processamento de linguagem natural (como, por exemplo, tokenização de palavras e sentenças, etiquetador gramatical, remoção de *stop words*, radicalização) para pré-processar o corpus com o objetivo de melhorar o processo de extração de informação.

- **Análise de Dados:** essa etapa é onde ocorre a extração de informação (análise do corpus). O objetivo dessa etapa é a extração de informação relevante do corpus através das regras e padrões definidos, utilizando para isso diversos métodos e enfoques.

- **Visualização:** nessa etapa as informações extraídas na etapa anterior são representadas de forma visual. A visualização das informações pode ser feita através de mecanismos simples (ex.: tabelas com as informações extraídas) ou complexos (ex.: interfaces 2D e 3D descrevendo as relações entre os conceitos extraídos).

- **Avaliação:** nessa etapa ocorre a avaliação das informações extraídas. Essa avaliação pode ser tanto de forma automática, utilizando-se, por exemplo, de técnicas de inteligência artificial, como manual, sendo feita por especialistas.

Juntamente com as abordagens e enfoques propostos, inúmeros mecanismos e ferramentas para análise do corpus (principalmente artigos científicos biomédicos) surgiram. Como destaque pode-se citar a ferramenta *SPECIALIST NLP Tools* (2010), desenvolvida pelo *The Lexical Systems Group* do *The Lister Hill National Center for Biomedical Communications* (COSTA, 2009). Essa ferramenta tem como objetivo ajudar desenvolvedores de aplicações biomédicas (relacionadas à UMLS (2010)) a analisar e processar linguisticamente o corpus (COSTA, 2009). Além de sua versão independente, o *SPECIALIST NLP Tools* também pode vir integrado ao MMTx (2010), uma aplicação, construída em Java, do MetaMap (2010) que tem como função mapear os termos encontrados no corpus para o UMLS *Metathesaurus* (2010) (COSTA, 2009).

No âmbito das ontologias da OBO, a ferramenta que mais se aproxima com o trabalho dessa dissertação é o *OBOL*. Mungall (2004), criador da ferramenta OBOL, defende a implantação de uma linguagem formal para descrição do campo definição e nome dos termos das ontologias da OBO, que possa ser compreendida por seres humanos e inferida por

máquinas. Para auxiliar nesse processo o OBOL é utilizado como uma ferramenta de manutenção da ontologia, verificando inconsistências, e como um meio de gerar definições computáveis (MUNGALL, 2004). Em seu artigo Mungall (2004) descreve a utilização dessa ferramenta, na ontologia *Gene Ontology,* com o objetivo de verificar inconsistências, através de uma análise linguística dos nomes de termos e das definições. Para a extração da relação do tipo *is_a,* por exemplo, Mungall (2004) utilizou três padrões baseados nos conceitos de *Genus-Differentia:*

G→ G*enus*

Q=(D)→ *Differentia*

D→ *Características*

1. G,Q = (D) ***is_a*** G,Q = (D') se D ***is_a*** D'

   Ex.: *chromoplast membrane* ***is_a*** *plastid membrane* [10], pois
   G→ *membrane*
   Q=(D)→ (*differentia = chromoplast*) → D= *chromoplast*,
   Q = (D') → (*differentia = plastid*) → D'= *plastid* e
   chromoplast ***is_a*** *plastid*
   Explicação: Dois termos podem possuir uma relação de *is_a* se os *genus* de ambos foram os mesmos e se há uma relação de *is_a* entre as características que diferenciam esses dois conceitos.

2. G,Q = (D) is_a G',Q = (D) se G is_a G'

   Ex.: *vitamin E biosynthesis* ***is_a*** *vitamin E metabolism*[10], pois
   G → *biosynthesis*
   G' → *metabolism*
   Q=(D) = → (*differentia = vitamin E*) e
   biosynthesis ***is_a*** *metabolism*
   Explicação: Dois termos que possuem *genus* diferentes e a mesma diferenciação podem possuir uma relação de *is_a* se os *genus* dos dois termos possuírem uma relação de *is_a*.

3. G,Q = (D) is_a G se ¬(□D' tal que D is_a D')

[10] Exemplo retirado do artigo de Mungall (2004).

Ex.: *primary septum* ***is_a*** *septum*[10], pois

G → *septum* e

Q=(D) → (*differentia = primary*)

Explicação: Se característica que diferenciam um conceito não for, e nem puder ser um conceito concreto de uma ontologia, então o conceito expresso com adição dessa característica deve possui um relação de *is_a* com seu *genus*.

Mais recentemente, o mecanismo OBOL foi aprimorado e está sendo usado como uma ferramenta do consórcio OBO para realizar o cruzamento entre ontologias da OBO, "produto cruzado", com o objetivo de completar o conhecimento dessas ontologias e encontrar erros e relações implícitas (MUNGALL *et al.*, 2010; GENE ONTOLOGY, 2010d). Entretanto, essa nova versão do OBOL está focando grande parte de seus esforços para procurar possíveis novas relações consideradas mais "modernas" (como, *occurs_in* e *adjacent_to*), em detrimento das relações mais básicas (como, *part_of* e *is_a*). Mungall *et al.* (2010) acreditam que essas novas relações mais modernas, se incorporadas às ontologias, poderão vir a completá-las, adicionando mais semântica e expressividade (MUNGALL *et al.*, 2010; GENE ONTOLOGY, 2010d).

Ainda nesse contexto, o trabalho de Pereira e Lôbo (2010) objetivou a criação de uma série de mecanismos e técnicas com base nos conceitos de processamento de linguagem natural para avaliação da *Gene Ontology*, com a finalidade de apontar possíveis automatizações e caminhos para reconhecimento de padrões nessa ontologia.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo descreveu de forma sucinta o que é extração de informação, mostrando sua diferença em relação à recuperação de informação, e também apresentou técnicas, métodos e enfoques utilizados nesse processo. Além disso, esse capítulo fez um estudo sobre mecanismos de extração de informação e análise linguística, fazendo uma comparação entre alguns deles e justificando a escolha nesse trabalho dos mecanismos GATE e NLTK.

Na seção 4.3 desse capítulo, foi feito um resumo sobre iniciativas de extração de informação no contexto biomédico, mostrando alguns trabalhos que serviram como base para a construção dessa dissertação.

A extração de informação permite que sejam explicitadas informações implícitas em corpus para diversas finalidades. No contexto desse trabalho são utilizados técnicas, ferramentas e enfoques de extração de informação para procurar e extrair possíveis informações implícitas nas ontologias biomédicas. Para isso é proposta, no próximo capítulo, uma abordagem para extração de informação no contexto das ontologias biomédicas divida em duas macroetapas. A primeira tem como finalidade conhecer o corpus das ontologias biomédicas e com isso permitir uma melhor precisão para a segunda macroetapa, que consiste na extração de informação propriamente dita. As informações extraídas através dessa abordagem podem vir a completar as ontologias e possivelmente melhorar os processos de reúso e anotação.

CAPÍTULO 5: ABORDAGEM PARA EXTRAÇÃO DE INFORMAÇÕES IMPLÍCITAS EM ONTOLOGIAS

Este capítulo descreve por macroetapas, etapas e subetapas a abordagem para extração de informações implícitas em ontologias, proposta nessa dissertação. No decorrer do desenvolvimento desse capítulo são mostrados os principais diferencias da abordagem e descritos com detalhes o que deve ser feito em cada passo de execução.

5.1 CONTEXTUALIZAÇÃO DA ABORDAGEM

O capítulo anterior descreveu alguns trabalhos relacionados ao tema dessa dissertação apresentando diversas abordagens, enfoques e ferramentas. Baseado no conhecimento adquirido com o estudo desses trabalhos e com a análise das ontologias biomédicas foi desenvolvida uma abordagem de extração de informação de ontologias e bases de dados que trabalha sobre a nomenclatura dos conceitos e sobre o campo definição. Como pontos positivos desta abordagem, citamos:

- Possui uma de macroetapa para conhecer o corpus, anterior ao processo de extração de informação propriamente dito, possibilitando com isso um provável ganho na etapa de extração de informação.

- Ser mais flexível, possibilitando mais interação com o usuário, que pode inclusive definir parâmetros durante a sua execução e receber dados preliminares através dos relatórios emitidos durante sua aplicação.

- Ser inicialmente focada, em cada ciclo de execução, em somente um tipo de relação ou objetivo para extração de informação, possuindo assim regras mais específicas e direcionadas.

- Ser aplicável a qualquer ontologia e não somente a uma específica.

- Ser facilmente adaptável a outros contextos e outras relações. Embora no contexto desse trabalho a abordagem seja aplicada a ontologias do domínio biomédico e para extração da relação de *is_a*, ela pode ser facilmente adaptada a outros domínios e bases de dados e para extração de qualquer outro tipo de relação ou objetivo.

- Utiliza-se de quatro enfoques para extração da informação, enquanto que a maioria das abordagens utiliza-se somente um enfoque.

A abordagem (esquematizada na Figura 14) proposta é divida em duas macroetapas: **Conhecer o Corpus** e **Trabalhar o Corpus**. A primeira macroetapa é divida em três etapas e tem por objetivo adquirir conhecimento sobre o corpus para aplicação desse conhecimento na segunda macroetapa. A segunda macroetapa é divida em duas etapas e tem por objetivo a extração e visualização de informações no corpus.

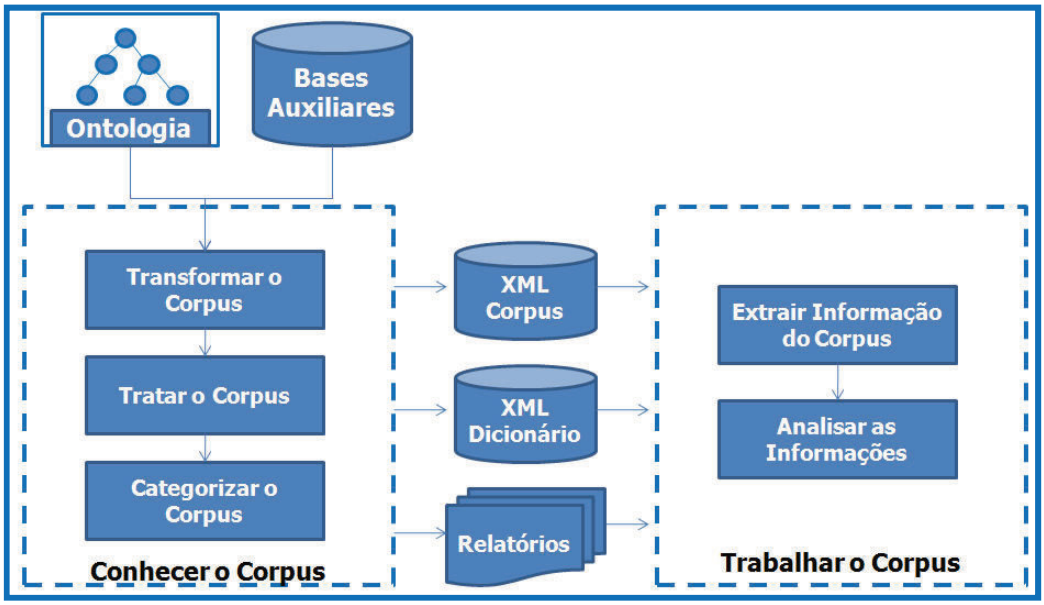

Figura 14: Esquema da Abordagem

5.2 CONHECER O CORPUS

O corpus utilizado nesse trabalho permite que ele seja estudado antes da extração de informação devido à abrangência desse corpus ser restrita a ontologias biomédicas. Ao se estudar o corpus é possível obter um ganho na extração de informação que poderá ser feita com maior precisão e eficácia. Desta forma, a macroetapa **Conhecer o Corpus** é divida em três etapas e serve para conhecer o corpus onde será aplicada a extração de informação propriamente dita. A primeira etapa **Transformar o Corpus** tem a função de transformar a ontologia e as bases auxiliares em um formato que será trabalhado pela abordagem. A segunda etapa, **Tratar o Corpus**, tem a função de analisar e marcar o corpus linguisticamente com a finalidade de facilitar a extração de informação. A terceira etapa, **Categorizar o**

Corpus, tem a função de fazer um estudo no corpus identificando padrões e conceitos relevantes.

5.2.1 TRANSFORMAR O CORPUS

A etapa de **Transformar o Corpus** (esquematizada na Figura 15) tem por objetivo recolher informações relevantes da ontologia pesquisada e das ontologias e base de dados auxiliares, transformando-as em dois arquivos XML que serão manipulados durante todo processo de extração de informação. O primeiro documento XML (XML CORPUS) contém somente os nomes dos termos, as relações, os sinônimos e as definições da ontologia estudada. O segundo documento XML (XML Dicionário) contém informações importantes que podem vir a completar o conhecimento extraído da ontologia como, por exemplo, sinônimos, termos e definições das ontologias e bases de dados auxiliares. A Figura 16 descreve um exemplo do padrão de XML entendido pela abordagem.

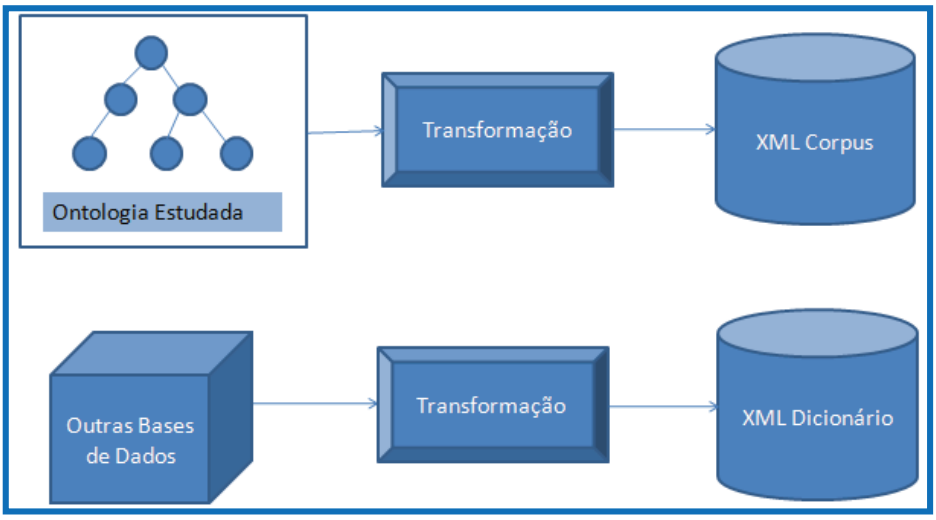

Figura 15: Esquema da Etapa de Transformar o Corpus

```
<termo id="GO_0015855">
<nome><![CDATA[pyrimidine transport]]></nome>
<definicao><![CDATA[The directed movement of pyrimidines, one of the two classes of
nitrogen-containing ring compounds found in DNA and RNA, into, out of or within a cell,
or between cells, by means of some agent such as a transporter or pore.]]></definicao>
<sinonimo><![CDATA[pyrimidine transmembrane transport]]></sinonimo>
<isa>GO_0015851</isa>
</termo>
```

Figura 16: Padrão do XML entendido pela abordagem

5.2.2 TRATAR O CORPUS

Após a preparação do corpus, o próximo passo é realizar o pré-processamento sobre ele. A etapa **Tratar o Corpus** (esquematizada na Figura 17) tem como objetivo marcar o corpus para facilitar as etapas posteriores. Essa etapa mantém as informações do XML Corpus e adiciona a ele outras informações como, por exemplo, classes gramaticais das palavras e sujeito das orações. Para auxiliar esse processo, nessa etapa da abordagem são utilizadas, em conjunto, os mecanismos GATE e NLTK para tratamento do corpus.

A Figura 18 ilustra as subetapas realizadas nessa etapa. As subetapas com linhas tracejadas são opcionais, uma vez que sua aplicação em alguns casos como, por exemplo, no corpus estudado nesse trabalho, não causam diferenças significativas, podendo até piorar o resultado da abordagem. Isso ocorre devido ao vocabulário do campo biomédico ser muito específico, contendo diversas particularidades. Entretanto essas etapas foram mantidas na abordagem, como opcionais, de modo a facilitar possíveis adaptações para extração de outros tipos de objetivos e o uso de outros tipos de base de dados.

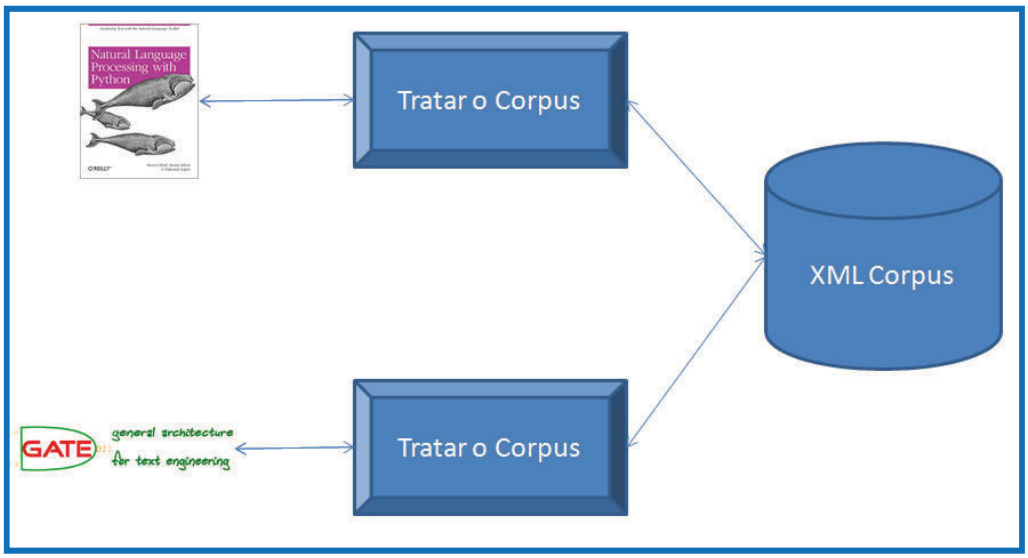

Figura 17: Esquema da etapa Tratar o Corpus

Figura 18: Subetapas da etapa Tratar Corpus

As subetapas da etapa de **Tratar o Corpus** são:

- **Tokenização de Palavras:** essa subetapa tem por objetivo dividir o corpus em tokens de palavras com algum significado. Essa tarefa pode ser muito complexa em ontologias biomédica, pois podem existir palavras separadas por números e letras gregas que somente possuem significado se utilizadas em conjunto.

- **Tokenização das Sentenças:** essa subetapa tem por objetivo separar todas as sentenças que formam o corpus. Essa divisão facilita a aplicação de algoritmos de analises linguística como etiquetador gramatical e análise sintática.

- **Radicalização (*Stemming*) (Opcional):** essa subetapa aplica o processo de radicalização às palavras que formam o corpus.

- **Lematização (Opcional):** essa subetapa aplica o processo de lematização às palavras que formam o corpus.

- **Remoção de *Stop Words* (Opcional):** essa subetapa retira do corpus os *Stop Words*, palavras com pouco significado semântico, ou seja, que se repetem muito no corpus como, por exemplo, artigos, numerais e preposições e etc.

- **Etiquetador Gramatical (*Part-of-speech tagging*):** essa subetapa faz anotações no corpus, relacionando cada palavra a sua classe gramatical como, por exemplo, advérbio, adjetivo e substantivo.

- **Análise Sintática (*Parsing*):** essa subetapa utiliza a tokenização das sentenças para construir árvores de *parsing* que possibilitem a anotação sintática do corpus de acordo com as classes sintáticas (sujeito, verbo e objeto e etc) de cada palavra ou conjunto de palavras.

- **Resolução de Problemas de Ambiguidade (Opcional):** essa subetapa tem como objetivo a retirada de ambiguidades do corpus, ou seja, palavras que possuem duplo sentido. Para isso, essa subetapa, avalia o contexto das palavras ou conjunto de palavras com duplo sentido, utilizando, por exemplo, de dicionários como o WordNet. Ex.: manga (fruta) e manga da camisa.

5.2.3 CATEGORIZAR O CORPUS

A etapa de **Categorizar o Corpus** divide-se em duas subetapas: **Calcular Relevância dos Elementos do Corpus**, onde é aplicado o enfoque baseado em estatística e **Descobrir Padrões no Corpus**, onde é utilizado o enfoque baseado em aprendizagem de máquina. O objetivo dessa etapa é permitir que o usuário da abordagem possua um entendimento global do corpus e possa com isso criar padrões mais direcionados para a extração de informação.

5.2.3.1 CALCULAR RELEVÂNCIA DOS ELEMENTOS DO CORPUS

A subetapa **Calcular Relevância dos Elementos do Corpus** (esquematizada na Figura 19) tem como objetivo identificar, por meio de um cálculo de relevância dos elementos do corpus, verbos que podem representar a relação que se deseja extrair e as palavras ou conjunto de palavras que podem vir a servir de *Stop Words* para o corpus. Por padrão, esse cálculo de relevância é feito utilizando o cálculo de frequência absoluta, mas pode ser trocado por qualquer outro cálculo estatístico. Esses cálculos de relevância são realizados da seguinte forma.

1. **Verbos Relevantes para o Domínio:** cria uma lista com todos os verbos do corpus mostrando a ocorrência de cada um.
2. ***Stop Words* para o Domínio[11]:** cria uma lista com todos os possíveis *Stop Words* do corpus mostrando a ocorrência de cada um. Isso é importante em contexto onde as ferramentas linguísticas não conseguem identificar com precisão as categorias gramaticais do corpus. A relação de *Stop Words* ocorre através da identificação das principais palavras que vem depois dos nomes dos termos conhecidos no corpus, permitindo que essa lista seja generalizada para os nomes não conhecidos.

Ao final da aplicação dessa subetapa as informações levantadas são apresentadas ao usuário em duas listas. A primeira lista auxilia o usuário a criar padrões para extração da relação pretendida. A segunda lista pode ser utilizada pela abordagem para identificação dos

[11] No contexto dessa abordagem são chamadas de *Stop Words* as palavras ou conjunto de palavras que permitem delimitar o término de um conceito em uma sentença.

principais *Stop Words* das bases pesquisadas, permitindo uma melhor precisão na extração de informação. Como produto dessa etapa o usuário faz a validação dessas listas com a finalidade de que essas informações possam ser utilizadas na macroetapa **Trabalhar o Corpus**.

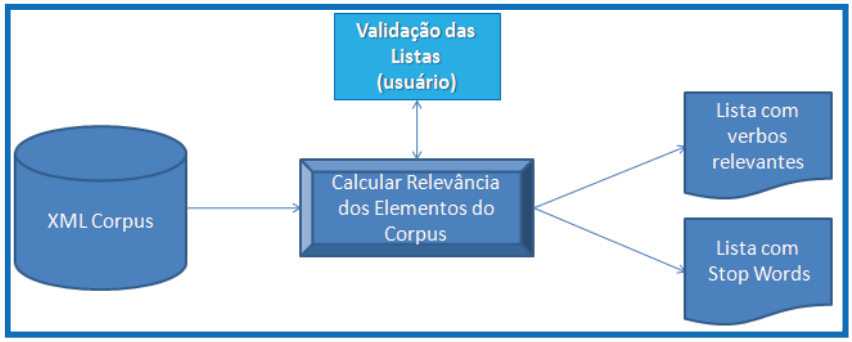

Figura 19: Esquema da subetapa Calcular Relevância dos Elementos do Corpus

5.2.3.2 DESCOBRIR PADRÕES NO CORPUS

A subetapa de **Descobrir Padrões no Corpus** (Figura 20) consiste em descobrir padrões de formação das estruturas do corpus através do enfoque baseado em aprendizagem de máquina.

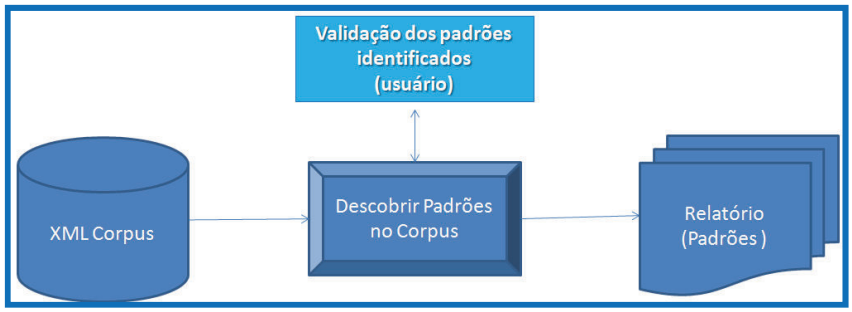

Figura 20: Esquema da subetapa Descobrir Padrões no Corpus

Para verificar a similaridade entre as definições e nomes dos termos, é utilizado um algoritmo similar à técnica de *Edit Distance,* que consiste na inserção, remoção ou troca de caracteres para que um dado conjunto de caracteres seja transformado em outro (CORMODE; MUTHUKRISHNAN, 2002). O algoritmo implementado na abordagem é treinado com um conjunto amostral, com resultado conhecido, de definições e nomes de termos das ontologias biomédicas, e serve para permitir maior eficiência na identificação dos padrões, pois a cada etapa do treinamento é feito um refinamento através da comparação do resultado conseguido

pelo algoritmo e do resultado esperado. Esse treinamento é importante para respeitar as particularidades de cada ontologia. A Figura 21 ilustra o processo de treinamento do algoritmo para descobrir padrões.

Posteriormente à aplicação desse algoritmo, as definições e os nomes dos termos são organizados em agrupamentos (clusters) com significados semelhantes. A noção de significado semelhante é calculada pela similaridade dada pelo algoritmo de *Edit Distance*. Através desse agrupamento é possível que o usuário identifique possíveis padrões no corpus e crie regras para serem utilizadas na macroetapa **Trabalhar o Corpus**.

Nessa etapa da abordagem o usuário também pode configurar o algoritmo utilizado para agrupar as definições e nomes dos termos com o intuito de identificar os padrões de definições já conhecidos e relevantes para o escopo da abordagem (Ex.: padrões de definições implementados pela GO, disponíveis no Anexo A).

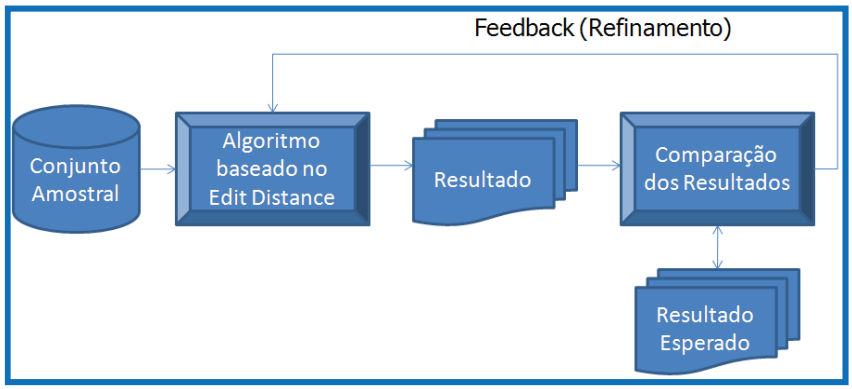

Figura 21: Treinamento do algoritmo baseado no Edit Distance

Ao final dessa subetapa é gerado um relatório contendo todos os padrões identificados e também as regras identificadas pelo usuário a partir desses padrões. Esse relatório é utilizado, na macroetapa seguinte, para servir como base na adaptação dos algoritmos utilizados pela abordagem.

5.3 TRABALHAR O CORPUS

A etapa de **Trabalhar o Corpus** é a macroetapa mais importante da abordagem, pois é nela que são extraídas as informações propriamente ditas. Essa macroetapa divide-se em duas etapas: **Extrair Informação do Corpus** e **Analisar as Informações Extraídas do Corpus**.

A etapa de **Extrair Informação do Corpus** tem por objetivo o preenchimento dos *templates* de saída que serão apresentados aos usuários. A etapa de **Analisar as Informações Extraídas do Corpus** tem por objetivo a visualização e análise dessas informações extraídas.

5.3.1 EXTRAIR INFORMAÇÃO DO CORPUS

A etapa **Extrair Informação do Corpus** tem por objetivo extrair conhecimento implícito do corpus, o qual pode incluir novos termos ou novas relações. Para isso ela se divide em três subetapas: **Criar Regras para o Corpus**, **Extrair Relações entre os Nomes dos Termos** e **Extrair Informações do Campo Definição**.

Nessa etapa da abordagem os algoritmos utilizados para extração de informação utilizam os enfoques baseados em regras e em dicionário. Ao final dessa etapa é gerado um relatório contendo todas as informações extraídas pela abordagem para validação do usuário ou especialista, na próxima etapa.

5.3.1.1 CRIAR REGRAS PARA O CORPUS

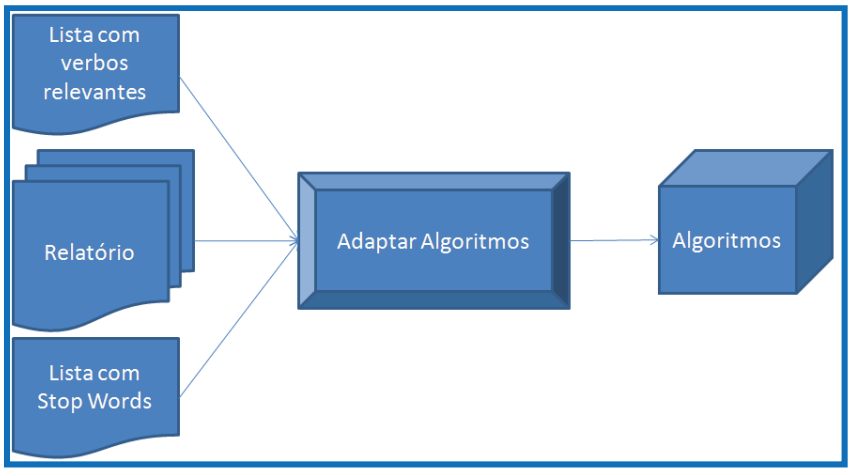

Figura 22: Esquema da subetapa Criar Regras para o Corpus

A subetapa **Criar Regras para o Corpus** (esquematizada na Figura 22) tem a função de adaptar os algoritmos (Algoritmos de Extração de Relação entre os Nomes e Algoritmos de Extração de Padrões da Definição) utilizados na abordagem para as especificidades do corpus,

de modo que eles trabalhem para encontrar os padrões identificados (descritos nos relatórios de padrões) na primeira macroetapa.

5.3.1.2 EXTRAIR RELAÇÕES ENTRE OS NOMES DOS TERMOS

A subetapa **Extrair Relações entre os Nomes dos Termos** (esquematizada na Figura 23) tem por finalidade extrair relações trabalhando os nomes dos termos e seus sinônimos. Nela são executados os algoritmos relacionados com a extração de relações através das nomenclaturas dos termos. Esses algoritmos extraem a informação pretendida pela abordagem através da análise dos nomes e sinônimos. Os algoritmos dessa subetapa são criados/adaptados na subetapa anterior utilizando os padrões identificados na primeira macroetapa. Essas mudanças nos algoritmos se fazem necessárias para que sejam respeitadas as particularidades de cada corpus avaliado.

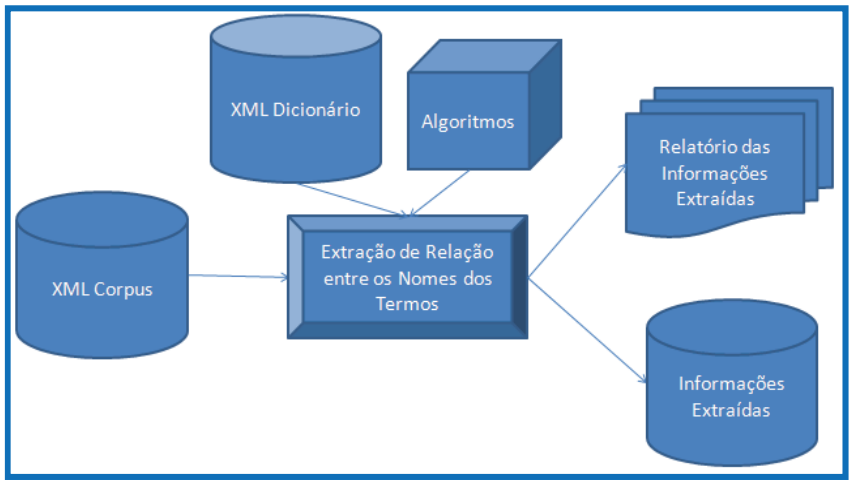

Figura 23: Esquema da subetapa Extrair Relações entre os Nomes dos Termos

Além das informações do corpus analisado (XML Corpus), essa etapa também utiliza o XML Dicionário com a finalidade de auxiliar no tratamento das informações extraídas. O XML Dicionário é acionado toda vez que for identificado, através dos nomes, um possível termo que não pertence à ontologia avaliada. Com a utilização desse XML é possível informar, juntamente com o novo termo encontrado, se ele pertence a alguma ontologia ou base de dados do XML Dicionário, permitindo com isso uma melhor análise do termo encontrado.

5.3.1.3 EXTRAIR INFORMAÇÕES DO CAMPO DEFINIÇÃO

A subetapa Extrair Informações do Campo Definição (esquematizada na Figura 24) tem por finalidade extrair relações e identificar novos termos através do campo definição. Para isso essa subetapa utiliza algoritmos voltados para trabalhar a definição implementados na subetapa **Criar Regras para o Corpus**.

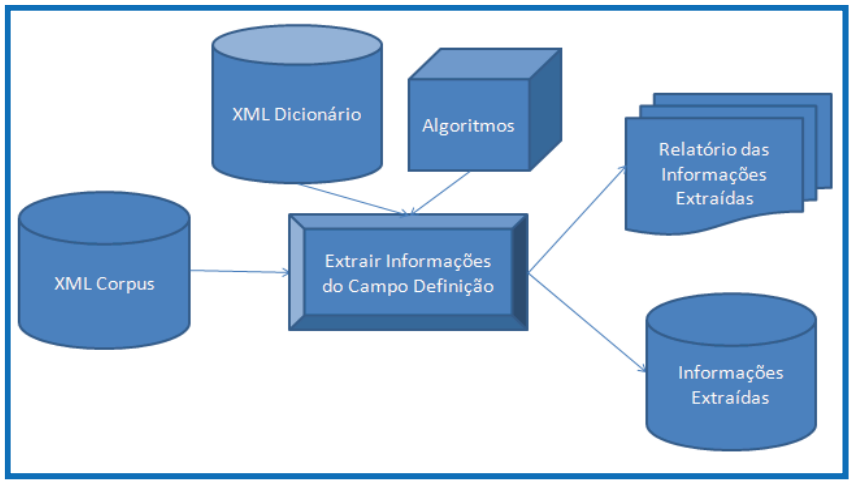

Figura 24: Esquema da subetapa Extrair Informações do Campo Definição

Assim como a subetapa anterior, essa subetapa também utiliza o XML Dicionário para identificar um termo referenciado na definição que, entretanto, não pertence à ontologia avaliada.

5.3.2 ANALISAR AS INFORMAÇÕES EXTRAÍDAS

A etapa de analisar as informações extraídas (Figura 25) tem por objetivo apresentar ao usuário da abordagem (que nesse caso pode ser um especialista) as informações extraídas nas etapas anteriores, com a finalidade de se fazer um estudo das informações como, por exemplo:

- Validação das informações encontradas.
- Análise da hierarquia dos termos com base nas novas relações (verificando a consonância com a hierarquia original da ontologia).

- Análise dos novos termos encontrados.
- Análise das novas relações.
- Análise dos benefícios das novas informações extraídas, como possíveis melhorias nos processos de reúso e anotação.
- Verificação de possíveis falhas nas ontologias.

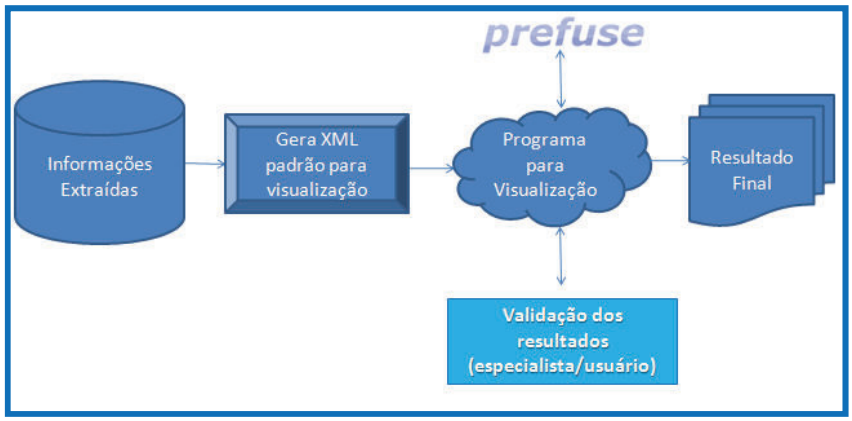

Figura 25: Esquema da etapa Analisar as Informações Extraídas

Para fazer essas análises os usuários utilizam os relatórios gerados nas etapas anteriores e também mecanismos de visualização construídos especificamente para o contexto desse trabalho como o EI-ONTO-Viewer (Figura 26).

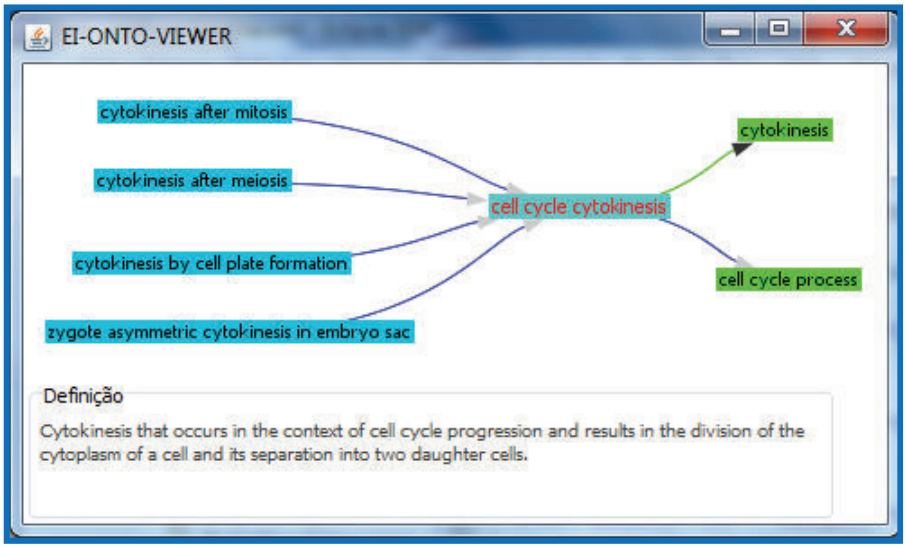

Figura 26: Ei-Onto-Viewer: mecanismo para visualização dos resultados da abordagem

O EI-ONTO-Viewer utiliza a biblioteca Prefuse (2010) para desenhar a ontologia como uma árvore, podendo atribuir cores diferentes as novas informações, permitindo ao usuário a identificação e validação visual das informações extraídas. Com isso, é possível, por

exemplo, verificar possíveis inconsistências e possíveis melhorias na hierarquia através da incorporação de novos termos e relações.

5.4 CONSIDERAÇÕES FINAIS

Este capítulo descreveu a abordagem para extração de informação em ontologias biomédicas. A abordagem proposta permite identificar informações implícitas nas ontologias biomédicas, com a finalidade de enriquecer o conteúdo desse tipo de ontologia. A idéia de utilização dessa abordagem é de que as informações implícitas nas ontologias poderiam torná-las mais expressivas, permitindo a verificação de inconsistências e falhas e, possivelmente, com isso, seja possível obter uma melhora nos processos de reúso e anotação.

Embora a abordagem proposta nessa dissertação seja voltada para ontologias e para a extração de relações implícitas, sendo aplicada no contexto desse trabalho às ontologias biomédicas e para extração de relações *is_a*, ela pode ser adaptada para receber qualquer tipo de base de dados e também para extrair qualquer tipo de relação ou objetivo. Para isso, conforme explicado nesse capítulo, a abordagem proposta foi dividida em duas macroetapas. A primeira é voltada para o estudo do corpus (no caso dessa abordagem, as ontologias biomédicas). Essa macroetapa tem a finalidade de conhecer o corpus e consequentemente proporcionar uma extração de informação mais direcionada, eficiente e eficaz. A segunda macroetapa consiste na extração de informação propriamente dita, e tem a finalidade de extração e validação de informações implícitas no corpus.

As etapas da abordagem foram propostas a partir das leituras bibliográficas e também da experimentação da tentativa e erro, sendo sempre ao final de cada tentativa refinada e aprimorada.

O próximo capítulo descreve com mais detalhes a implementação dessa abordagem e a avaliação experimental realizada em duas ontologias biomédicas, que testou sua aplicação e eficácia.

CAPÍTULO 6: AVALIAÇÃO EXPERIMENTAL

Este capítulo descreve a avaliação experimental realizada para avaliar a abordagem. Nessa avaliação experimental a abordagem proposta foi aplicada a duas ontologias, do consórcio OBO, com a finalidade de extrair informações implícitas. Ao final da aplicação dessa abordagem as ontologias utilizadas foram alinhadas com a finalidade de verificar se houve uma possível melhora, nesse processo, através da incorporação das novas relações. Além da avaliação experimental, esse capítulo também descreve o mecanismo EI-ONTO que implementa a abordagem.

6.1 CONTEXTUALIZAÇÃO DO EXPERIMENTO

O reúso de ontologias é um processo importante, uma vez que permite a utilização de diversas ontologias em conjunto, aumentando o conhecimento contido em cada uma delas individualmente (Campos, 2009a).

A abordagem de extração de informação proposta nesta dissertação tem como objetivo a extração de conhecimentos implícitos de quaisquer ontologias ou base de dados, através de estruturas como o campo definição ou a nomenclatura dos termos com vista a aumentar a qualidade e semântica das ontologias ou base de dados. No contexto desse trabalho essa abordagem foi aplicada para a extração de novos termos e relações hierárquicas do tipo é_um (*is_a*) explorando, para isso, tanto o campo definição como a nomenclatura dos termos.

Para testar essa abordagem, nesse capítulo é proposto um experimento que tem por objetivo verificar o possível ganho conseguido através da utilização dessas novas relações e termos nas ontologias biomédicas escolhidas. Para isso, o experimento desse capítulo foi planejado em duas fases. A primeira fase tem por objetivo a aplicação da abordagem em duas ontologias do consórcio OBO com o intuito de recuperar as informações implícitas, verificando a quantidade e qualidade das informações extraídas. A segunda fase tem por objetivo a realização de dois alinhamentos: o primeiro com as ontologias, sem nenhuma informação adicional e o segundo alinhamento utilizando as ontologias com as novas relações, tendo como resultado os pares obtidos em cada um dos alinhamentos.

6.2 PLANEJAMENTO DO EXPERIMENTO

Conforme descrito anteriormente, o experimento realizado nessa dissertação foi divido em duas fases. Na primeira fase a abordagem foi aplicada em duas ontologias biomédicas do consórcio OBO. Ao final desse processo as informações extraídas foram analisadas e validadas permitindo já um primeiro resultado. Na segunda fase as informações das possíveis novas relações validadas na primeira etapa foram acrescentadas às ontologias (esquematizado na Figura 27) e foram feitos dois alinhamentos para permitir uma comparação do alinhamento dessas ontologias antes e depois da aplicação da abordagem (esquematizada na Figura 28). Isso foi feito com a finalidade de verificar uma possível melhoria nesse processo, através da utilização das informações implícitas.

Para fazer esse alinhamento foi escolhida a ferramenta FOAM (EHRIG; SURE, 2005), devido a essa ferramenta ter sido identificada no trabalho de Silva (2010) como uma das ferramentas mais viáveis para alinhamento de ontologias.

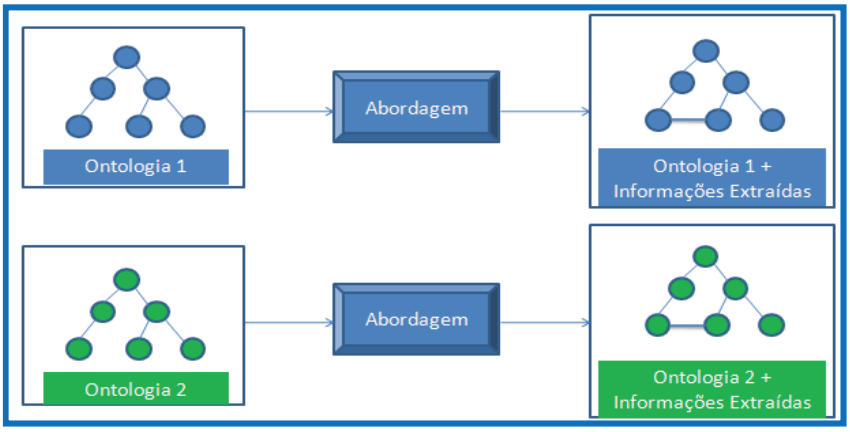

Figura 27: Esquema da primeira fase do experimento

Na ferramenta FOAM foram utilizados os seguintes parâmetros na execução do alinhamento entre as ontologias:

- Alinhamento: Totalmente Automático.
- Número de iterações: 10.

- Valor de corte[12]: 0,97.

- Estratégia: Arvore de Decisão (*Decision Tree*).

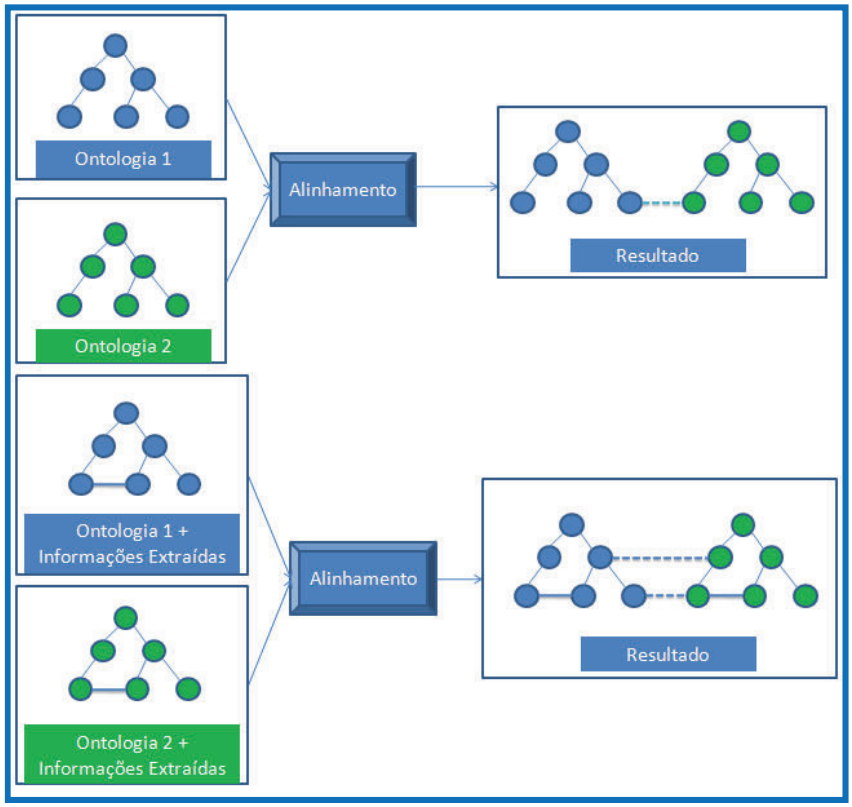

Figura 28: Esquema da segunda fase do experimento

Depois de definidas todas as fases da avaliação experimental, partiu-se para a escolha das ontologias que seriam utilizadas. A primeira ontologia escolhida foi a GO, pois, conforme explicado em capítulos anteriores, essa ontologia é a mais utilizada atualmente, sendo a principal ontologia do consórcio OBO. Depois da escolha dessa ontologia, partiu-se para a escolha da segunda ontologia que seria utilizada nessa avaliação experimental. As ontologias biomédicas do consórcio OBO testadas nesse processo foram:

- *INOH Molecular Role.*

- *INOH Event.*

- *Brenda.*

[12] Valor mínimo para que um par de termos seja considerado como resultado do alinhamento, ou seja, somente pares com valor de corte maiores ou iguais ao valor definido são considerados no resultado final.

- *Pathway*.

Essas ontologias foram testadas por apresentarem conceitos e temática com alguma similaridade com a GO e por já terem sido objeto de estudo no decorrer dessa dissertação. Entretanto, dentre essas ontologias, a ontologia escolhida foi a *INOH Event*, por quatro motivos:

1. Essas duas ontologias já foram alinhadas por outro trabalho (SILVA, 2010) do grupo de pesquisa do qual essa dissertação faz parte.

2. A *Gene Ontology* possui alinhamento manual com essa ontologia. Isso foi percebido através da análise de uma lista de 800 termos da *Gene Ontology* utilizados no sequenciamento do organismo *Trypanosoma Rangeli* do trabalho de Wagner (2006) onde foi verificada uma coincidência verbal de 26 termos com os termos da *INOH Event* (SILVA, 2010).

3. A *INOH Event* e a *Gene Ontology* foram objetos de estudo, no grupo ONTOTAXO[13], pelas Professoras Maria Luiza Almeida Campos e Hagar Espanha Gomes que verificaram nessas ontologias diversas possibilidades de extração terminológicas. Esses testes foram conduzidos nos moldes do trabalho de Campos (2009b).

4. A *INOH Event*, assim como a *Gene Ontology*, possui uma interface de visualização web[14] para busca e navegação na hierarquia dos termos, facilitando com isso o processo de validação dos experimentos (SILVA, 2010).

Além disso, através das pesquisas do grupo GRECO[15], do qual essa dissertação faz parte, foi verificado que dos 26 termos com coincidência verbal com a GO, descritos na Tabela 8, adaptada de Silva (2010), 25 pertenciam ao ramo *Biological Process* da GO e somente 1 ("*binding*") pertencia ao ramo *Molecular Function* (SILVA, 2010). A partir dessa informação optou-se, de modo a restringir o trabalho computacional e a validação manual do trabalho por somente trabalhar com o ramo *Biological Process* da *Gene Ontology*, conforme já havia sido feito na avaliação experimental de Silva (2010).

[13] www.ontotaxo.uff.br
[14] AmiGO: <http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>
INOH Ontology Viewer: <http://www.inoh.org:8083/ontology-viewer>
[15] www.greco.ppgi.ufrj.br

As ontologias utilizadas nesse experimento foram obtidas através do site da OBO no dia 22 de novembro de 2010 às 14h30min. A Tabela 9 mostra as informações sobre as ontologias utilizadas nessa avaliação experimental.

Tabela 8: Termos com coincidência verbal presentes na GO e na *INOH Event*

Termo	Classe GO	Classe INOH
actin cytoskeleton organization	GO_0030036	IEV_0001323
binding	GO_0005488	IEV_0000004
BMP signaling pathway	GO_0030509	IEV_0000863
cell adhesion	GO_0007155	IEV_0000105
cell differentiation	GO_0030154	IEV_0000002
cell growth	GO_0016049	IEV_0000091
cell motility	GO_0048870	IEV_0000103
cell proliferation	GO_0008283	IEV_0000076
cell-cell signaling	GO_0007267	IEV_0002611
cytokinesis	GO_0000910	IEV_0000079
cytoskeleton organization	GO_0007010	IEV_0000092
DNA repair	GO_0006281	IEV_0000872
DNA replication	GO_0006260	IEV_0000393
embryo development	GO_0009790	IEV_0000397
immune response	GO_0006955	IEV_0000098
lamellipodium assembly	GO_0030032	IEV_0000093
MAPKKK cascade	GO_0000165	IEV_0000445
nervous system development	GO_0007399	IEV_0002710
organ morphogenesis	GO_0009887	IEV_0000576
oxidative phosphorylation	GO_0006119	IEV_0001431
phosphorylation	GO_0016310	IEV_0000005
regulation of transcription	GO_0045449	IEV_0001833
response to stress	GO_0006950	IEV_0001337
response to wounding	GO_0009611	IEV_0001341
regulated secretory pathway	GO_0045055	IEV_0002539
translation	GO_0006412	IEV_0000219

Fonte: adaptado de Silva (2010)

Tabela 9: Informações sobre as ontologias utilizadas

Informação	*Biological Process*	*INOH Event*
Termos	20392	3775
Termos obsoletos	519	149
Termos com definição (retirando os obsoletos)	19870	522

6.3 APLICAÇÃO DA ABORDAGEM

Nessa seção será descrita a aplicação da abordagem nas duas ontologias utilizadas nessa avaliação experimental utilizando-se, para isso, as etapas da abordagem descritas no capítulo 5. Além disso, essa seção também descreve com mais detalhes o mecanismo que implementa essa abordagem, mostrando algumas telas utilizadas durante sua aplicação.

6.3.1 CONHECER O CORPUS

6.3.1.1 TRANSFORMAR O CORPUS

Na aplicação dessa etapa, a ontologia testada e as bases auxiliares foram transformadas em dois arquivos XML (XML Corpus e XML Dicionário). Esses arquivos XML somente continham o nome dos termos, suas relações, seus sinônimos e sua definição. Além disso, na transformação para esses arquivos XML também foram eliminados os termos obsoletos.

Tabela 10: Conteúdo do XML Corpus e do XML Dicionário

XML Corpus	XML Dicionário
Biological Process (GO)	*INOH Event* *INOH Molecular Role* *Cellular Component* *Molecular Function* *Pathway* *Brenda*
INOH Event	*Biological Process* *INOH Molecular Role* *Cellular Component* *Molecular Function* *Pathway* *Brenda*

Para esse experimento foram escolhidas apenas 6 ontologias para compor o XML Dicionário (Tabela 10). Essas ontologias foram escolhidas por possuírem características e/ou termos em comum com as ontologias estudas e também por elas já terem sido trabalhadas no estudo dessa dissertação.

A Figura 29 ilustra a interface utilizada para carregar e transformar as ontologias e bases de dados auxiliares. Nessa parte do mecanismo é possível inserir as ontologias e as bases de dados auxiliares e também a ontologia principal que será tratada, para que sejam transformadas no XML padrão entendido pelo mecanismo EI-ONTO.

Figura 29: Interface para carregar e transformar a ontologia e bases de dados auxiliares

6.3.1.2 TRATAR O CORPUS

Nessa etapa da abordagem foram adicionadas informações ao XML Corpus, utilizando recursos como etiquetador gramatical e análise sintática, importantes para a aplicação das demais etapas. As subetapas opcionais dessa etapa não foram aplicadas nesse experimento, pois em testes anteriores verificou-se que no contexto desse trabalho, essas etapas atrapalhariam a extração de informação, ou que não se faziam necessárias.

Uma das dificuldades encontradas nessa etapa foram alguns erros gerados pelas ferramentas de análise linguística, que não conseguiram identificar algumas categorias gramaticais devido à complexidade de escrita do corpus biomédico, que em suas definições possuem várias frases com sujeito oculto e verbos substantivados como, por exemplo, "*Combining with interleukin-10 to initiate a change in cell activity*".

6.3.1.3 CATEGORIZAR O CORPUS

A etapa de categorizar o corpus foi onde ocorreu o estudo do corpus propriamente dito. Primeiro, foi aplicada a subetapa **Calcular Relevância dos Elementos do Corpus**. Na aplicação dessa subetapa foram identificados verbos relevantes que poderiam descrever a relação alvo desse trabalho (*is_a*) em cada uma das ontologias, conforme mostrado na Tabela 11. Além disso, devido ao problema de que o corpus biomédico possui dificuldade em ser marcado por ferramentas linguísticas, nessa subetapa também foram identificados os *Stop Words*[16] mais frequentes dessas ontologias para que fossem guardados e utilizados como pontos de parada (*breakpoint*) no processo de extração de conceitos. Essas palavras foram validadas pelo usuário da abordagem e estão descritos na Tabela 12.

Tabela 11: Verbos que caracterizam a relação *is_a* nas Ontologias Biomédicas

is a – na definição	*such as* – na definição (possível relação de hiponímia)

Tabela 12: *Stop Words* mais frequentes das Ontologias Biomédicas

adjacent	*resulting*
after	*that*
before	*through*
comprising	*what*
containing	*whatever*
found	*where*
from	*whereby*
induced	*whether*
involved	*which*
located	*who*
occurring	*whose*
proceeding	*with*
result	*within*

Depois dessa subetapa foi aplicada a subetapa **Descobrir Padrões no Corpus** que identificou através do algoritmo baseado na técnica *Edit Distance*, padrões na definição e nos nomes dos termos para que pudessem ser utilizados no processo de extração de informação.

Esses padrões foram levantados de acordo com o resultado da aplicação desse algoritmo juntamente com a observação da formação das relações e nomes dos termos das ontologias pesquisadas. Os padrões identificados através da nomenclatura dos termos foram:

[16] No contexto dessa abordagem são chamados de *Stop Words* as palavras ou conjunto de palavras que permitem delimitar o término de um conceito em uma sentença.

(1) **Padrão:** *Substring*

Descrição: Se um termo é uma *substring* final de outro, ele deve possivelmente possuir uma relação hierárquica com esse termo ou com algum sinônimo exato desse termo.

Observação: A parte diferente dessa *substring* deve possuir somente substantivo e adjetivos, pois caso possua outras classes gramaticais isso pode indicar outro tipo de relação como no exemplo: "**Termo 1:** *establishment of localization* (GO:0051234) não é *is_a localization* (GO:0051179)". Onde não existe uma relação de *is_a* e sim uma de *part_of*.

Exemplos:

***Biological Process*:** *alkane transport* (GO:0015895) *is_a transport* (GO:0006810)

***INOH Event*:** *cyclohexemide medication* (IEV:0001569) *is_a medication* (IEV:0000292)

(2) **Padrão:** Regulation

Descrição: Todos os termos que têm *positive* ou *negative regulation of [x]* devem possivelmente possuir em sua hierarquia uma relação com o termo *regulation of [x]* ou algum sinônimo de *regulation of [x]*.

Observação: No *Biological Process* da GO todas as relações de *positive regulation of [x]* e *negative regulation of [x]* possuem como sinônimos exatos *up regulation of [x]*, *upregulation of [x]*, *up-regulation of [x] (positive)* e *down regulation of [x]*, *downregulation of [x]*, *down-regulation of [x] (negative)*. Esses sinônimos podem ser ignorados uma vez que representam os mesmos conceitos descritos pelo *positive* e *negative regulation*.

Exemplos:

***Biological Process*:** *positive regulation of viral reproduction* (GO:0048524) *is_a regulation of viral reproduction* (GO:0050792)

***INOH Event*:** *negative regulation of ras-raf-mapk signaling by spred* (IEV:0003544) *is_a regulation of ras-raf-mapk signaling by spred* (IEV:0003543)

(3) **Padrão:** Padrão do OF

Descrição: Todos os termos que possuem "*of*" e "*in*" devem possivelmente possuir em sua hierarquia uma relação com o termo que antecede essas preposições.

Exemplos:

***Biological Process*:** *localization of cell* (GO:0051674) *is_a localization* (GO:0051179)

***INOH Event*:** *nuclear import of stat dimer* (IEV:0000048) *is_a nuclear import* (IEV:0000018)

Já os padrões identificados através do campo definição foram:

(4) **Padrão:** Primeira Definição

Descrição: Todos os termos que possuem uma definição que comece com "A", "An", "The", "Any ou "Those" devem, possivelmente, possuir em sua hierarquia uma relação com o primeiro termo expresso na definição após essas palavras.

Exemplos:

***Biological Process*:**

Termo: *actin filament reorganization involved in cell cycle* (GO:0030037)

Definição: *The **cell cycle process** whereby rearrangement of the spatial distribution of actin filaments and associated proteins occurs*.

Relação encontrada: *actin filament reorganization involved in cell cycle* (GO:0030037) *is_a cell cycle process* (GO:0022402)

***INOH Event*:**

Termo: *nonribosomal peptide biosynthesis* (IEV:0003805)

Definição: *The **biosynthetic process** whereby peptide bond formation occurs in the absence of the translational machinery. Examples include the synthesis of antibiotic peptides, and glutathione*.

Relação encontrada: *nonribosomal peptide biosynthesis* (IEV:0003805) *is_a biosynthetic process* (Não pertencente a *INOH Event*)

(5) Padrão: Segunda Definição

Descrição: Todas as vezes que for encontrada "*is a*" no campo definição, os termos anteriores e posteriores a esse conjunto de palavras possivelmente possuem uma relação hierárquica.

Observação: Embora seja considerado um erro conceitual a definição de um termo dentro de outro, nas ontologias pesquisadas diversas definições possuem essa característica, inclusive incentivadas pelas próprias definições padronizadas da *Gene Ontology* disponíveis no Anexo A desse trabalho. A extração dessa informação pode ser feita utilizando algoritmos parecidos com o da regra da primeira definição.

Exemplos:

***Biological Process*:**

Termo: *regulation of necroptosis* (GO:0060544)

Definição: *Any process that modulates the rate frequency or extent of necroptosis. **Necroptosis** is a **necrotic cell death** process that results from the activation of endogenous cellular processes, such as signaling involving death domain receptors and Toll-like receptors*.

Relação encontrada: *necroptosis* (GO:0070266) *is_a necrotic cell death* (GO:0070265)

***INOH Event*:**

Termo: *open tracheal system development* (IEV:0003791)

Definição: *The process whose specific outcome is the progression of an open tracheal system over time, from its formation to the mature structure. An **open***

> ***tracheal system*** *is a* ***respiratory system****, a branched network of epithelial tubes that supplies oxygen to target tissues via spiracles. An example of this is found in Drosophila melanogaster.*
>
> **Relação encontrada:** *open tracheal system* (Não pertecence a *INOH Event*) *is_a respiratory system* (Não pertence a *INOH Event*)

A Figura 30 ilustra a interface de interação com o usuário para identificação dos padrões. Nessa interface, os usuários podem escolher o nível de similaridade[17] e o número mínimo de repetições que uma estrutura deve ter para ser considerado um padrão. Além disso, também é possível, nessa parte do mecanismo, acrescentar um arquivo XML de padrões já identificados (como, por exemplo, os da GO, descritos no Anexo A desse trabalho) para que os mesmos sejam encontrados no corpus.

Figura 30: Interface para identificação de padrões

Além dos padrões identificados com a aplicação da primeira macroetapa também foi observado que os padrões do OBOL definidos por Mungall (2004) poderiam ser utilizados para extração de relações da nomenclatura dos termos e que os padrões de Hearst (1992; 1998) poderiam auxiliar a extração de informação do campo definição. Os padrões do OBOL e de Hearst são detalhados a seguir.

(6) Padrão: OBOL - Regra 1

[17] O nível de similaridade é determinado pela semelhança entre dois conjuntos de caracteres (palavras) através do algoritmo implementado na abordagem (baseado no *Edit distance*).

Descrição: *G→ Genus*

Q=(D)→ Differentia

D→ Características

G, Q = (D) ***is_a*** G, Q = (D') se D ***is_a*** D'

Um termo é *is_a* de um segundo termo se o *genus* dos dois termos foram o mesmo e se a *differentia* do primeiro tiver uma relação de *is_a* com a *differentia* do segundo.

Observação: A explicação mais detalha sobre esse padrão pode ser encontrada na seção 4.3 desta dissertação.

Exemplos:

***Biological Process*:** *catabolism by host of symbiont xylan* (GO:0052365) *is_a metabolism by host of symbiont xylan* (GO:0052420)

***INOH Event*:** *reduction in cytosol* (IEV:0000249) is_a *oxidation/reduction in cytosol* (IEV:0001059)

(7) **Padrão:** OBOL - Regra 2

Descrição: *G→ Genus*

Q=(D)→ Differentia

D→ Características

G, Q = (D) ***is_a*** G', Q = (D) se G ***is_a*** G'

Um termo é *is_a* de um segundo termo se a *differentia* dos dois termos foram a mesma e se o *genus* do primeiro tiver uma relação de *is_a* com o *genus* do segundo.

Observação: A explicação mais detalha sobre esse padrão pode ser encontrada na seção 4.3 desta dissertação.

Exemplos:

***Biological Process*:** *o-glycoside catabolic process* (GO:0016142) *is_a o-glycoside metabolic process* (GO:0016140)

***INOH Event*:** Não foi possível encontrar, através do padrão do OBOL - Regra 2, nenhum par de termos que pudesse representar uma relação de *is_a* correta na *INOH Event*.

(8) **Padrão:** OBOL - Regra 3

Descrição: *G→ Genus*

Q=(D)→ Differentia

D→ Características

G, Q = (D) ***is_a*** G se ¬(∃D' tal que D ***is_a*** D')

Um termo é *is_a* de seu *Genus* se sua *differentia* for um qualificador para o termo, ou seja, a *differentia* não pode ser um conceito que é expresso ou que pode ser expresso na ontologia.

Observação: A explicação mais detalha sobre esse padrão pode ser encontrada na seção 4.3 desta dissertação. Nessa dissertação esse padrão foi absorvido pelo padrão da *Substring*.

Exemplos:

***Biological Process*:** *primary cell wall biogenesis* (GO:0009833) *is_a cell wall biogenesis* (GO:0042546)

***INOH Event*:** *smooth muscle cell differentiation* (IEV:0001824) *is_a muscle cell differentiation* (IEV:0001823)

(9) **Padrão:** Padrão de Hearst

Descrição: Nesse padrão são aplicadas às frases do campo definição as 6 regras identificadas por Hearst (1992; 1998), com a finalidade de encontrar relações de hiponímia, que podem descrever as relações de *is_a.*

1. NP_0 *such as* NP_1 {,NP_2 ... , (*or* | *and*) NP_i}, onde para cada i>0 tem-se um Hipônimo (NP_i,NP_0).
2. *such* NP *as {*NP_1 *,}* {(or | and)} NP*$_i$, onde para cada NP_i tem-se um Hipônimo (NP_i,NP).
3. NP_1 {, NP_i}* {,} *or other* NP, onde para cada NP_i tem-se um Hipônimo (NP_i,NP).
4. NP_1 {, NP_i}* {,} *and other* NP, onde para cada NP_i tem-se um Hipônimo (NP_i,NP).
5. NP {,} *including* {NP_0 ,}* {*or* | *and*} NP_i, onde para cada NP_i tem-se um Hipônimo (NP_i,NP).
6. NP {,} *especially* {NP_0 ,}* {*or* | *and*} NP_i, onde para cada NP_i tem-se um Hipônimo (NP_i,NP).

Observação: Mais detalhes na seção 3.4.2 desta dissertação.

Exemplos:

***Biological Process*:**

Termo: *cellular response to corticosteroid stimulus* (GO:0071384)

Definição: *A change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a corticosteroid hormone stimulus. A corticosteroid is a steroid hormone that is produced in the adrenal cortex. Corticosteroids are involved in a wide range of physiologic systems* ***such as*** *stress response, immune response and regulation of inflammation, carbohydrate metabolism, protein catabolism, blood electrolyte levels, and behavior. They include glucocorticoids and mineralocorticoids.*

Relação encontrada:

Hipônimo (*immune response*, *physiologic system*).

immune response (GO:0006955) *is_a physiologic system* (Não pretence a GO)

***INOH Event*:**

Termo: *cytoplasm organization* (IEV:0001321)

Definição: *A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of the cytoplasm. The cytoplasm is all of the contents of a cell excluding the plasma membrane and nucleus, but* ***including*** *other subcellular structures.*

Relação encontrada:

Hipônimo (*plasma membrane* , *subcellular structures*).

plasma membrane (Não pretence a *INOH Event*) *is_a subcellular structures* (Não pretence a *INOH Event*)

Nessa etapa da abordagem o padrão 8 (OBOL - Regra 3) foi descartado devido a esse padrão ser coberto pelo padrão 1 (*Substring*). Outras considerações importantes na aplicação dessa parte da abordagem são que, embora essa etapa tenha sido aplicada separadamente em cada uma das ontologias, foram observados praticamente os mesmos padrões de formação nessas ontologias tanto na primeira quanto na segunda subetapa, e, por esse motivo optou-se por descrever em conjunto esses padrões e informações levantadas. Além disso, também foi observado nessas ontologias que a maior parte das palavras grifadas em letras maiúsculas, tanto nos nomes dos termos como nos sinônimos e definições, são abreviações de nomes e que poderiam ser tratadas de maneira especial.

Outro ponto importante verificado nessa parte da abordagem foi a análise das definições padronizadas da GO (Anexo A), para com isso se tentar extrair mais algum tipo de padrão onde fosse possível a extração da relação *is_a*. Entretanto, após essa análise, verificou-se que as possíveis regras para extração da relação *is_a* descobertas através dessas definições já eram atendidas pelos padrões identificados.

6.3.2 TRABALHAR O CORPUS

6.3.2.1 EXTRAIR INFORMAÇÃO DO CORPUS

A etapa de **Extrair Informação do Corpus** é onde ocorre a extração de dados propriamente dita. Nessa etapa da abordagem foram adaptados os algoritmos (subetapa **Criar**

Regras para o Corpus) e as informações foram extraídas (subetapa **Extrair Relações entre os Nomes dos Termos** e subetapa **Extrair Informações do Campo Definição**).

Na subetapa de **Criar Regras para o Corpus** os padrões identificados na macroetapa anterior foram implementadas no mecanismo EI-ONTO. Além disso, nessa etapa também foram tratadas algumas abreviações da ontologia (como, por exemplo, *BMP receptor signaling pathway* (GO:0030509), onde BMP significa *bone morphogenetic protein*). O significado dessas abreviações foi descoberto por meio dos sinônimos dos termos da própria ontologia ou das bases auxiliares que possuíam essas abreviações como sinônimos exatos.

É importante ressaltar que todos os algoritmos implementados (algoritmos relacionados ao campo definição e a nomenclatura dos termos) nessa subetapa utilizaram os sinônimos exatos da ontologia para extração e validação das informações extraídas. Com a utilização dos sinônimos foi possível a extração de mais informação e também a retirada de erros na extração que seriam inseridos caso só se considerasse os nomes dos termos das ontologias. Um exemplo desse tipo de erro seria a verificação de uma relação ou termo que pelo nome parecia inexistente, mas que é expresso através dos sinônimos. Por exemplo, não existe nenhum termo com o nome de *morphogenesis*. Entretanto, *anatomical structure morphgenesis* (GO:0009653) possui como sinônimo exato o nome *morphogenesis*. Isso impede que a relação *"morphogenesis of a branching epithelium" is_a "morphogenesis"* seja marcada como inexistente, pois *"morphogenesis of a branching epithelium*" (GO:0061138) *is_a "anatomical structure morphgenesis*" (GO:0009653).

A Tabela 13 faz um resumo dos padrões implementados nessa subetapa e mostra quais tipos de informação cada padrão recupera.

Legenda:

primeiro termo *is_a* segundo termo

2PO – o par de termos encontrados pertencem à ontologia pesquisada, ou seja, esse tipo encontra uma relação entre dois termos existentes na ontologia.

1PO2NO – somente o primeiro termo encontrado pertence à ontologia pesquisada, ou seja, esse tipo encontra uma relação entre um termo pertencente à ontologia e outro que é citado na ontologia, nesse caso como um pai do termo pertencente à ontologia.

1NO2PO - somente o segundo termo encontrado pertence à ontologia pesquisada, ou seja, esse tipo encontra uma relação entre um termo pertencente à ontologia e outro que é citado na ontologia, nesse caso como um filho do termo pertencente à ontologia.

2NO – nenhum dos dois termos pertence à ontologia pesquisada, ou seja, esse tipo encontra uma relação entre dois termos que não pertencem à ontologia pesquisada, porém são citados nessa ontologia.

Tabela 13: Resumo dos padrões de extração implementados

Padrão	Foco	2PO	1O2NO	1NO2O	2NO
Substring	Nome dos Termos	X			
Regulation	Nome dos Termos	X	X		
Regra do OF	Nome dos Termos	X	X		
OBOL-Regra 1	Nome dos Termos	X			
OBOL-Regra 2	Nome dos Termos	X			
Primeira Definição	Campo Definição	X	X		
Segunda Definição	Campo Definição	X	X	X	X
Padrão de Hearst	Campo Definição	X	X	X	X

Na subetapa **Extrair Relações entre os Nomes dos Termos** foram utilizados os padrões para extração de informação com foco na nomenclatura, com a finalidade de extração de informação aproveitando-se dos nomes dos termos e sinônimos. A Tabela 14 descreve a quantidade de relações extraídas em cada padrão na aplicação dessa subetapa.

Tabela 14: Quantidade de relações extraídas em cada padrão relacionado à nomenclatura dos termos

Padrão	Tipo	*INOH Event*-Quantidade	BP-Quantidade
Substring	2PO	31	436
Regulation	2PO	0	4
Regulation	1PO2NO	82	652
Padrão do OF	2PO	29	15
Padrão do OF	1PO2NO	1447	1227
OBOL Regra 1	2PO	45	7
OBOL Regra 2	2PO	1	216

Na subetapa **Extrair Informação do Campo Definição** foram utilizados os padrões para extração de informação com foco no campo definição com a finalidade de extração de informação dessa estrutura. A Tabela 15 descreve a quantidade de relações extraídas em cada padrão com a aplicação dessa subetapa.

Tabela 15: Quantidade de relações extraídas em cada padrão relacionado ao campo definição

Padrão	Tipo	*INOH Event*-Quantidade	BP-Quantidade
Primeira Definição	2PO	2	55
Primeira Definição	1PO2NO	313	10821
Segunda Definição	2PO	0	0
Segunda Definição	1PO2NO	6	45
Segunda Definição	1NO2PO	0	2
Segunda Definição	2NO	3	383
Padrão de Hearst	2PO	0	0
Padrão de Hearst	1PO2NO	0	17
Padrão de Hearst	1NO2PO	0	0
Padrão de Hearst	2NO	38	490

6.3.2.2 ANALISAR AS INFORMAÇÕES EXTRAÍDAS

Na etapa de **Analisar as Informações Extraídas**, as informações extraídas na etapa anterior foram analisadas para verificação de sua consistência e relevância. Ao final dessa etapa foi feito um levantamento para quantificar e qualificar os resultados das possíveis novas relações e dos possíveis novos termos extraídos pela abordagem. A Tabela 16 descreve o resultado da validação feita com auxílio de biólogos e também com a correlação das informações extraídas com as informações expressas nas ontologias biomédicas (como as informações expressas no campo definição e na hierarquia dos termos). Algumas observações que merecem destaque nessa subetapa do experimento são descritas na Tabela 17.

Tabela 16: Validação preliminar dos resultados

Padrão	Tipo	*INOH Event* validadas/total	BP validadas/total
Substring	2PO	21/31	311/436
Regulation	2PO	0/0	4/4
Regulation	1PO2NO	82/82	652/652
Padrão do OF	2PO	29/29	14/15
Padrão do OF	1PO2NO	1394/1447	571/1227
OBOL Regra 1	2PO	45/45	2/7
OBOL Regra 2	2PO	0/1	125/216
Primeira Definição	2PO	2/2	55/55
Primeira Definição	1PO2NO	306/313	10558/10821
Segunda Definição	2PO	0/0	0/0
Segunda Definição	1PO2NO	6/6	37/45
Segunda Definição	1NO2PO	0/0	2/2
Segunda Definição	2NO	3/3	244/383
Padrão de Hearst	2PO	0/0	0/0
Padrão de Hearst	1PO2NO	0/0	12/17
Padrão de Hearst	1NO2PO	0/0	0/0
Padrão de Hearst	2NO	2/38	37/490

Através da aplicação dessa abordagem verificou-se a omissão de termos e relações que possivelmente deveriam pertencer à estrutura hierárquica da ontologia, pois aparecem na definição de outros termos ou têm relação com os nomes dos termos. Essa inconsistência pode representar problemas na interpretação de um termo ou eventuais falhas na estrutura classificatória originalmente criada, dificultando com isso o reúso da ontologia.

Tabela 17: Observações sobre os padrões adotados na abordagem

Padrão	Observação
Todos	- Embora tratada pela abordagem a abreviação dos nomes ainda foi um fator de dificuldade para o processo de extração de informação dessas ontologias. - Na *INOH Event* menos de 15% dos termos tem definição, o que dificulta o processo de validação das informações extraídas nessa ontologia. - Durante todo o desenvolvimento dessa dissertação se percebeu uma evolução constante da GO, o que não se verificou com tanta frequência na *INOH Event*.
Substring	- O Padrão *Substring* recupera muitas relações de *part_of*, mesmo tendo sido incluída no algoritmo uma etapa que visa retirar preliminarmente esse tipo de relação. - Na aplicação desse padrão é preciso um cuidado adicional, pois nem toda *substring* pode representar uma relação. Termos que são compostos químicos, por exemplo, precisam de uma avaliação preliminar de biólogos.
OBOL	- Percebe-se que grande parte dos termos da ontologia usa os padrões OBOL na formação de sua estrutura hierárquica. Possivelmente isso se deve ao fato desse padrão ter sido descrito pelo próprio consórcio da OBO. - Os padrões do OBOL trabalham melhor com o ramo componente celular da GO, uma vez que esse ramo descreve características mais concretas do domínio biomédico (componentes celulares), diferentemente dos ramos processo biológico e função molecular que descrevem processos e funções respectivamente.
OF	- O número de termos recuperados pelo padrão do OF foi um pouco elevado devido ao algoritmo estar configurado para recuperar o *Genus* de todos os termos. Alternativamente esse algoritmo poderia recuperar somente os Genus dos termos que fossem descritos por alguma das ontologias ou bases auxiliares, evitando a ocorrência de termos vagos.
Regulation	- Nessa abordagem, para adicionar mais semântica aos termos, optou-se por relacionar todo *negative e postive regulation* a um *regulation,* embora, seja sabido, que na estrutura hierárquica da GO e na *INOH Event* isso não ocorra com tanta frequência, pois nela um *negative regulation* pode ser filho de outro *negative regulation* sem se relacionar diretamente com nenhum *regulation*.
Primeira Definição	- O algoritmo utilizado para trabalhar a primeira definição encontrou diversos termos que são usados para descrever o termo dono da definição, porém não existem formalmente na ontologia trabalhada e sim em outras ontologias. Isso demonstra claramente um "*gap*" do conhecimento representado nessas ontologias.
Segunda Definição	- Grande parte dos termos que são definidos dentro de outro não pertencem à ontologia. - Embora seja condenada pelos autores que estudam as definições, a prática de utilizar uma definição dentro de outra é bem comum nas ontologias biomédicas, sendo útil para a descrição de termos inexistentes nessas ontologias.
Padrão de Hearst	- Na maior parte dos casos do padrão de Hearst na ontologia, os nomes encontrados não representam classes (termos) e, sim, instâncias.

Uma segunda questão que complementa essa primeira é a de que é possível o enriquecimento do conhecimento das ontologias através da incorporação de novas relações e termos encontrados através da abordagem. Porém, não podemos afirmar que isto deva sempre

resultar em um benefício para as ontologias, pois a relevância dessas novas informações encontradas deve ser avaliada caso a caso para evitar um aumento excessivo de complexidade da ontologia provocando um maior trabalho para os curadores. Além disso, outro ponto que merece destaque é que sempre que se for adicionada uma relação ou um novo termo à ontologia é preciso ter cautela, olhar com cuidado a árvore hierárquica dos termos evitando com isso que se liguem termos genéricos demais com termos específicos demais, fazendo com que a ontologia perca um pouco de sua semântica, obtendo um resultado inverso ao esperado. Com isso, conclui-se que nem sempre uma relação aparentemente óbvia pode ser importante para uma ontologia.

Através da análise do experimento, além dos resultados propriamente ditos, foram constatadas nas ontologias diversas outras questões ou problemas que poderiam ser objeto de estudo de trabalhos futuros como:

- Nos próprios nomes dos termos podem ser encontrados outros tipos de relações que poderão ser usadas para enriquecer as ontologias futuramente. Ex.: *small gtpase mediated signal transduction* onde podemos encontrar a relação *"mediated_by"*. De fato, isso já está sendo objeto de estudo em uma iniciativa proposta pelo trabalho de Mungall *et al.* (2010) que está tentando adicionar novas relações mais modernas como, por exemplo, *occurs_in,* às ontologias da GO, para que seja possível garantir mais semântica.

- Nas ontologias biomédicas ocorrem diversos homônimos e abreviações que dificultam a validação das informações extraídas. Ex.: *pancreatic b cell differentiation*.

- Existem estruturas de termos que são construídas de tal forma que permite que um termo seja transitivamente *is_a* e *part_of* de um termo superior, fazendo que nas inferências das ontologias essas falhas fiquem evidentes.

- Na GO ocorrem mudanças constantes como remoção de termos e mudanças nos nomes dos termos e nas estruturas. Entretanto, algumas vezes, em vez de tornar o termo obsoleto e criar um novo, é aproveitado o mesmo identificador do termo antigo, contrariando algumas regras do consórcio OBO.

- O campo definição, embora cada vez mais padronização, ainda possui falhas em sua construção, contrariando os princípios da definição defendidos por Dahlberg (1978; 1981) e outros autores. Pode-se citar como exemplo o fato de ainda existirem definições circulares, definições que explicitam mais de um termo em seu corpo e também definições que dificultam uma inferência de conhecimento por máquinas como, por exemplo, as que utilizam verbos substantivados e não possuem uma estrutura clara de sujeito, verbo e predicado. Ex.: O termo *system process* (GO:0003008) é definido dentro da definição do termo *modulation by symbiont of host system process* (GO:0044055) cuja a definição é: *The alteration by a symbiont organism of the functioning of a system process in the host. A* ***system process*** *is a multicellular organismal process carried out by any of the organs or tissues in an organ system*.

6.4 ALINHAMENTOS

6.4.1 PREPARAÇÃO PARA OS ALINHAMENTOS

Conforme descrito anteriormente, essa fase do experimento foi utilizada para testar a relevância das informações extraídas. Para isso foram feitos dois alinhamentos, ambos utilizando a ferramenta FOAM e os parâmetros[18] descritos anteriormente. O primeiro alinhamento foi realizado com as ontologias originais e o segundo com as ontologias acrescidas das novas informações implícitas encontradas através da abordagem em cada um dos focos (foco na nomenclatura dos termos e foco nas definições). Embora o mecanismo que implementa a abordagem tenha uma funcionalidade de analisar a hierarquia[19] e escolher a melhor posição para encaixar um possível novo termo, nesse experimento optou-se por somente acrescentar as possíveis novas relações, não acrescentando os possíveis novos termos por dois motivos:

1. Mesmo com a ajuda do mecanismo que implementa a abordagem, a inclusão de um novo termo é extremante complexa e requer uma análise minuciosa da

[18] Alinhamento: Totalmente Automático; Número de Iterações: 10; Valor de Corte: 0,97; estratégia: Árvore de Decisão *(Decision Tree)*.

[19] Esse algoritmo utiliza como critério de avaliação a similaridade entre as nomenclaturas dos termos, a estrutura hierárquica do termo (caso o termo pertença a outra ontologia), a estrutura hierárquica do local onde o termo será inserido e também outras medidas como quantidade de pais e filhos.

hierarquia.

2. Optou-se por não aumentar a quantidade de termos das ontologias do segundo alinhamento, fazendo com que elas fiquem com a mesma quantidade de termos das ontologias do primeiro alinhamento.

Além disso, devido ao ramo *Biological Process* da *Gene Ontology* por si só ser uma ontologia de grande porte, optou-se por dividi-la em segmentos que seriam alinhados com a *INOH Event*. Isso foi feito com a finalidade de restringir os resultados para que os mesmos fossem mais facilmente validados por um especialista. Dessa forma, utilizou-se uma estratégia similar a empregada no do trabalho de Silva (2010), utilizando-se o *Galen Segmenter*[20], e fragmentando a ontologia em 25 fragmentos com base nos termos da Tabela 8 pertencentes ao ramo *Biological Process*.

Nessa etapa do experimento foi preciso adicionar as novas relações às ontologias, com isso foram adicionadas 510 relações (existia uma relação que foi encontrada por dois padrões) na BP e 97 relações na *INOH Event*. Essas novas relações foram adicionadas como novos recursos através da utilização do JENA[21].

Com a finalidade de garantir a convergência dos resultados do alinhamento optou-se nesse experimento por alinhar os pares de ontologias cinco vezes cada, considerando ao final somente os resultados que apareciam em pelo menos três vezes. Isso se fez necessário, pois o algoritmo utilizado para alinhamento no FOAM é baseado em árvore de decisão e em alguns casos pode ser não determinístico. Ao final dos alinhamentos foram encontrados 57 pares no primeiro e 60 pares no segundo alinhamento.

6.4.2 ANÁLISE DOS ALINHAMENTOS

Depois dos processos de alinhamento, os resultados foram enviados para dois biólogos, com experiência em sequenciamento e anotação genômica, para que pudessem ser validados. Essa validação ocorreu nos moldes do trabalho de Silva (2010), obedecendo inclusive à classificação criada nesse trabalho.

[20] http://www.co-ode.org/galen
[21] API Java utilizada para navegação e manipulação de ontologias. Fonte: <http://jena.sourceforge.net >

Nessa classificação os biólogos avaliam os resultados dividindo-os em cinco categorias, conforme mostrado na Tabela 18. Além dos resultados totalmente corretos e totalmente incorretos nessa avaliação é possível ter resultados parciais que podem ser úteis em determinados contextos onde o anotador não está somente interessado na relação de equivalência, mas sim em outros tipos de relações como "é_um (*is_a*)" e "parte_de (*part_of*)". Dessa forma os resultados intermediários embora, não representem uma relação de equivalência, permitem agregar informações ao alinhamento (SILVA, 2010). A Tabela 19 e as Figuras 31 e 32 apresentam os resultados da validação para cada um dos alinhamentos[22].

Tabela 18: Classificação dos Alinhamentos

Grau	Significado
5	Correto
4	Relação Forte
3	Relação Média
2	Relação Fraca
1	Incorreto

Fonte: Silva (2010)

Tabela 19: Resultados dos Alinhamentos

Classificação	Resultados			
	Alinhamento1		Alinhamento2	
Grau	Quantidade	Porcentagem(%)	Quantidade	Porcentagem(%)
5	39	68%	47	78%
4	4	7%	3	5%
3	7	12%	7	12%
2	6	11%	2	3%
1	1	2%	1	2%

Através desse experimento foi possível evidenciar que a utilização das novas relações às ontologias permitiu uma melhoria no processo de alinhamento realizado, pois no segundo alinhamento houve um aumento na qualidade e na quantidade de pares alinhados. Analisando os resultados pode-se verificar um aumento de 8 pares corretamente alinhados (aumento de 10%) e também uma diminuição dos pares com relações fracas no segundo alinhamento, diminuição de 8%. Além disso, também foi possível verificar que o acréscimo das novas relações às ontologias fez com que houvesse mais possibilidades de caminhos e mais informações disponíveis durante a análise, permitindo aos algoritmos da ferramenta maior poder de decisão e escolhas mais precisas.

Desta forma, podemos concluir, com esse experimento, que o acréscimo das novas

[22] Nos pares de alinhamento onde houve divergência no grau atribuído pelos biólogos, optou-se por utilizar o menor grau entre as duas classificações.

relações implícitas nas ontologias biomédicas evidenciou uma possível melhoria no processo de alinhamento dessas ontologias aumentando o número de pares equivalentes e diminuindo o número de pares com relação fraca.

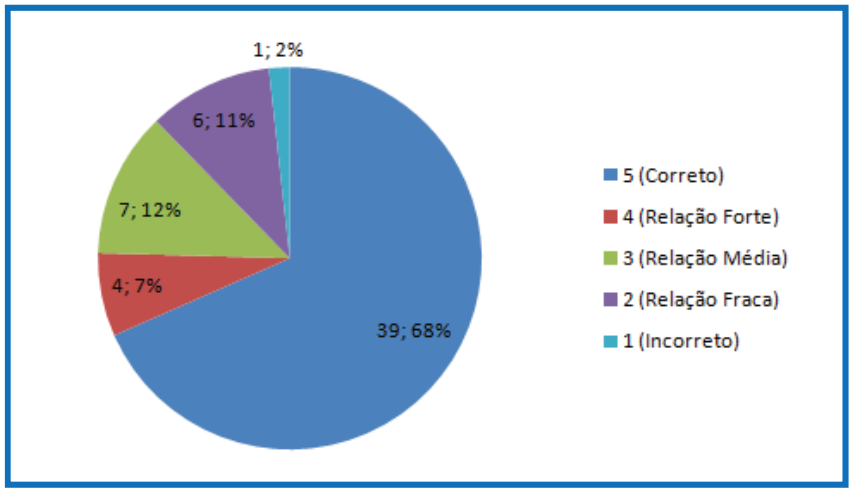

Figura 31: Gráfico do resultado do primeiro alinhamento

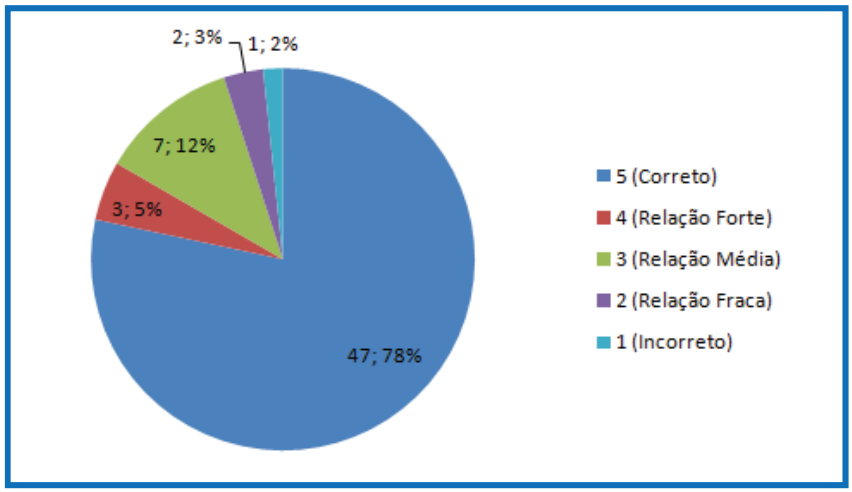

Figura 32: Gráfico do resultado do segundo alinhamento

6.5 CONSIDERAÇÕES FINAIS

Esse capítulo apresentou a avaliação experimental realizada no âmbito dessa dissertação. A avaliação experimental ocorreu em duas fases. Na primeira fase houve a aplicação da abordagem a duas ontologias biomédicas do consórcio OBO e na segunda fase foi feita uma comparação dos alinhamentos das ontologias com e sem as relações extraídas na primeira fase.

Através do experimento foi possível verificar dois resultados principais que validam a abordagem. O primeiro resultado consiste na quantidade e qualidade das informações extraídas da ontologia, evidenciando a existência de informações implícitas e mostrando que é possível resgatá-las. O segundo resultado é evidenciado através de uma possível melhoria no processo do alinhamento com o acréscimo das novas relações. Isso demonstra que as novas relações podem trazer benefícios para a ontologia como aumento da expressividade, melhorando os processos de reúso e de anotação.

Além desses resultados, também foi possível verificar problemas e questões na ontologia que merecem ser estudados mais a fundo como: estudo e maior padronização das definições para que sejam entendidas por seres humanos e computáveis por máquinas, e também a extração de novos tipos de relações.

É importante frisar que essa dissertação está em consonância com os estudos atuais da OBO para melhoramento das ontologias. Prova disso é que durante o desenvolvimento deste trabalho foram observados diversos esforços por parte da OBO para melhorias dessas ontologias, como aumento do número de termos definidos, melhoria nos mecanismo de inferência e também incorporação de novas relações na hierarquia.

CAPÍTULO 7: CONCLUSÃO

Este capítulo descreve as considerações finais desse trabalho. Primeiramente são descrito o cenário da pesquisa e os antecedentes que motivaram essa dissertação. Depois disso são enumeradas as principais contribuições, trabalhos futuros e limitações.

A pesquisa desenvolvida nessa dissertação teve inicio em 2007 a partir de trabalhos realizados nas ontologias biomédicas durante a iniciação científica. Durante esses trabalhos e através de leituras bibliográficas foram constatados, nas ontologias biomédicas, diversos problemas estruturais e semânticos ocasionados pela clara omissão de termos e relações.

Contudo, também foi verificado que nas ontologias biomédicas existiam muitas informações implícitas que poderiam complementar o conhecimento expresso. Além disso, também se observou uma visível melhoria dessas ontologias com o passar do tempo, com destaque para a *Gene Ontology*. Nessa ontologia foi possível identificar constantes melhorias e incorporação de informações implícitas, além de estudos para incorporação de novas relações como no trabalho de Mungall *et. al.* (2010). Outro exemplo de iniciativas desse tipo são as melhorias das definições da GO, que no decorrer desse trabalho foram padronizadas e conseguiram obter 100% dos termos definidos.

Através dessas observações e com a finalidade de aumentar a qualidade das ontologias, corrigindo alguns dos problemas constatados, foi proposta como produto dessa dissertação uma abordagem para trabalhar as ontologias biomédicas com o intuito de enriquecimento do conhecimento nelas embutido. Vale destacar com isso a relevância desse trabalho, que mostrou estar alinhado com diversas outras iniciativas do consórcio OBO para melhoria das ontologias biomédicas.

No desenvolvimento da abordagem foram estudados diversos enfoques, técnicas e outras abordagens para extração de informação. A abordagem desenvolvida nesse trabalho foi construída na base da tentativa e do erro, sendo sempre ao final de cada aplicação refinada e aprimorada fazendo uma espécie de refinamento (*feedback*) até se chegar à abordagem final descrita nessa dissertação. Uma preocupação constante na construção da abordagem foi o desenvolvimento de etapas genéricas, permitindo que a abordagem seja adaptada para qualquer contexto de extração.

A avaliação experimental dessa abordagem ocorreu com a sua aplicação a duas ontologias do consórcio OBO, obtendo dois resultados principais. O primeiro resultado foi conseguido através da extração e validação de centenas de termos e relações implícitas nessas ontologias. O segundo resultado foi obtido através do processo de alinhamento, onde foi evidenciada uma melhoria, desse processo, através do acréscimo das novas relações, aumentando tanto a qualidade como a quantidade de pares alinhados, demonstrando que as informações extraídas podem ser utilizadas para melhoria no processo de reúso. Os resultados dessa avaliação experimental serviram para comprovar os benefícios da utilização da abordagem e também para levantar questões que merecem serem objetos de pesquisa em trabalhos futuros. Dentre estas questões, está à aplicação da abordagem através do mecanismo EI-ONTO a outras ontologias e a outros contextos, como já vem sendo feito (CAMPOS *et al.,* 2010), obtendo resultados parciais satisfatórios.

7.1 CONTRIBUIÇÕES

Essa dissertação deixa como principal contribuição a criação da abordagem e sua implementação através do mecanismo EI-ONTO. Além disso, a pesquisa dessa dissertação confirmou diversos questionamentos relacionados às pesquisas biomédicas e também levantou diversos aspectos que merecem ser estudados com mais detalhes. Destacam-se os padrões e delimitadores de conceitos descobertos, que podem ser aplicados a diversas outras ontologias biomédicas.

O objetivo da abordagem proposta nessa dissertação era de construí-la com etapas genéricas com a finalidade de poder ser adaptada a diversos contextos, entretanto podendo suprir as particularidades de extração de informação no contexto biomédico. Os principais diferencias da abordagem em relação às demais estão ligados diretamente ao fato dela apresentar uma macroetapa de conhecer o corpus, anterior ao processo de extração propriamente dito, fazendo com que seja possível analisar o corpus, previamente, e consequentemente aumentar a eficiência do processo de extração de informação como um todo. Além disso, a utilização de quatro enfoques (dicionário, regras, aprendizagem de máquina e estatístico) e a possibilidade de maior interação com o usuário permite uma melhor extração de informação, a identificação de falhas mais cedo e também uma melhoria nos algoritmos da abordagem.

Outra contribuição desse trabalho foi o desenvolvimento do mecanismo EI-ONTO, o qual implementa essa abordagem, e também diversos outros pequenos programas que poderão auxiliar futuros usuários da abordagem. Entre esses programas incluem-se programa de visualização das ontologias, de troca de id por *label* e programas estatísticos para análises da saída dos algoritmos e para comparação de ontologias. Esses programas são importantes, pois permitem a automatização de estudos e pesquisas, principalmente da área de Ciência da Informação, onde questões semânticas de corpus terminológicos vêm sendo estudadas há muito tempo. A implementação da abordagem e dos programas auxiliares visa permitir, por exemplo, automatizar alguns estudos que antes eram feitos manualmente, como os do trabalho de Campos (2009b).

Além disso, a aplicação da abordagem, os vários testes realizados durante o seu desenvolvimento e o estudo da literatura serviram para comprovar dois fatos. O primeiro deles é que existem informações implícitas nas ontologias biomédicas e que se essas informações forem trabalhadas corretamente podem servir para enriquecer essas ontologias. O segundo fato é que o tema estudado nessa dissertação também é uma grande preocupação do consórcio OBO, que tem várias iniciativas com esse propósito. De fato, testes realizados nesse trabalho em versões mais antigas e mais novas da ontologia, apontam uma constante evolução das ontologias, que se verificou na padronização dos nomes dos termos e da definição e também na incorporação de relações anteriormente implícitas. Somado a isso, também foram verificados, em artigos científicos, estudos para enriquecimento das ontologias biomédicas através da incorporação de novos tipos de relação.

7.2 TRABALHOS FUTUROS E LIMITAÇÕES

Essa dissertação teve como principal objetivo pesquisar as informações implícitas nas ontologias biomédicas e proporcionar conhecimento para servir de base para diversos outros caminhos de pesquisa. Um dos possíveis trabalhos futuros dessa dissertação é a aplicação dessa abordagem em sua versão final para extração de informação em outros ramos da *Gene Ontology* e em outras ontologias do consórcio OBO e também em outros contextos como, por exemplo, em artigos científicos e bases de dados. Além disso, sugere-se a aplicação da abordagem com a finalidade de extração de outros tipos de relações como, por exemplo, *part_of*, *adjacent_to* ou *occurs_in*. Isso poderia envolver estudos mais aprofundados sobre a

relevância e o *tradeoff* de incorporar essas novas informações às ontologias. Essa extração de novas relações e novos termos poderia se aproveitar das definições padronizadas da GO para geração de suas definições, em um trabalho em consonância com o que está sendo desenvolvido na OBO por Mungall *et al.* (2010).

Outros trabalhos possíveis são o aperfeiçoamento dos algoritmos implementados na abordagem, como, por exemplo, a implementação do processamento de linguagem natural em Prolog[23] e aprofundamento das técnicas para incorporação dos novos termos à hierarquia das ontologias. Além disso, pode-se desenvolver interfaces mais intuitivas para o mecanismo EI-ONTO no estilo *drag-and-drop* e também implementar algoritmos para auxílio na documentação como, por exemplo, algoritmos para identificação de palavras-chaves com a finalidade de verificação automática das temáticas das ontologias.

Outro aspecto que merece ser estudado com mais detalhe é a verificação dos princípios definidos pela OBO e dos próprios princípios ontológicos na construção e manutenção de ontologias. Pois os maiores problemas encontrados, devem-se à não utilização desses princípios, levando a diversos erros como, por exemplo, o emprego de relações equivocadas e omissão clara de termos e relações.

Outra sugestão de trabalhos futuros, adaptando a sugestão de Silva (2010), seria a união das informações dos alinhamentos no FOAM conseguidos pela abordagem tradicional, pela abordagem proposta nessa dissertação e pela abordagem proposta no trabalho de Silva (2010) e utilizá-las em um contexto real de anotação, permitindo a união dessas três fontes de informações e, com isso, um maior poder de escolha ao anotador. Isso poderia ser feito através de interfaces interativa como a implementada nessa dissertação através do mecanismo EI-ONTO e utilizada na validação das informações.

[23] http://www.swi-prolog.org/

REFERÊNCIAS

AGARWAL, P.; SEARLS, D.B. **Literature Mining in Support of Drug Discovery.** Briefings in Bioinformatics, v.9, n.6, p.479-492, 2008.

ALMEIDA, A.C.B. **BIOANOT: Um Sistema Multi-Agentes para Notificação de (Re) Anotações de Sequências em Bancos De Dados Genômicos.** 140 f. Dissertação de Mestrado (Mestrado em Sistemas e Computação). Instituto Militar de Engenharia, 2006.

ALMEIDA, M.B.; BAX, M.P. **Uma Visão Geral sobre Ontologias: Pesquisa sobre Definições, Tipos, Aplicações, Métodos de Avaliação e de Construção.** Ciência da Informação, Brasília, v.32, n.3, 2003.

ALTSCHUL, S.F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, D.J. **Basic Local Alignment Search Tool.** Journal of Molecular Biology, v.215, n.3, p.403-410, 1990.

ANANIADOU, S.; McNAUGHT, J. **Text Mining for Biology and Biomedicine.** Norwood: Artech House, 286p., 2006.

BELLOZE, K.T. **Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular.** 148 f. Dissertação de Mestrado (Mestrado em Sistemas e Computação). Instituto Militar de Engenharia, 2007.

BENSON, D.A.; KARSCH-MIZRACHI, I.; LIPMAN, D.J.; OSTELL, J.; WHEELER, D.L. **GenBank: Update.** Nucleic Acids Research, v.32, p.23-26, 2004.

BERARDINI, T.Z.; MUNDODI, S.; REISER, L.; HUALA, E.; GARCIA-HERNANDEZ, M.; ZHANG, P.; MUELLER, L.A.; YOON, J.; DOYLE, A.; LANDER, G.; MOSEYKO, N.; YOO, D.; XU, I.; ZOECKLER, B.; MONTOYA, M.; MILLER, N.; WEEMS, D.; RHEE, S.Y. **Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies.** Plant Physiology, v.135, p.745-755, 2004.

BERKELEY BIOINFORMATICS OPEN SOURCE PROJECTS. Disponível em: <www.berkeleybop.org> Acesso em: 10 ago. 2010.

BIOMEDICAL INFORMATICS RESEARCH NETWORK. Disponível em: <www.birncommunity.org> Acesso em: 10 dez. 2010.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. O'Reilly Media. 2009. Disponível em: <http://www.nltk.org/book> Acesso em: 20 fev. 2010.

BODENREIDER, O.; BURGUN, A. **Biomedical Ontologies.** In: Medical Informatics Knowledge Management and Data Mining in Biomedicine. CHEN, H.; FULLER, S.S.; FRIEDMAN, C.; HERSHP, W. (eds.) Medical Informatics Knowledge Management and Data Mining in Biomedicine. New York: Springer-Verlag, USA, p.211-236, 2005.

BONTCHEVA, K.; TABLAN, V.; MAYNARD, D.; CUNNINGHAM, H. **Evolving GATE to Meet New Challenges in Language Engineering.** Natural Language Engineering, v.10, p.349-373, 2004.

BONTCHEVA, K.; MAYNARD, D.; TABLAN, V.; CUNNINGHAM, H. **GATE: A Unicode-based Infrastructure Supporting Multilingual Information Extraction.** In: Proceedings of Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages, Bulgaria, 2003.

BODENREIDER, O.; STEVENS, R. **Bio-ontologies: Current Trends and Future Directions.** Briefings in Bioinformatics, v.7, n.3, p.256-274, 2006.

BRACHMAN, R.J. **What IS-A Is and Isn't: An Analysis of Taxonomic Links in Semantic Networks.** Computer. v.16, n.10, p.30-36, 1983.

BRENDA TISSUE/ENZYME SOURCE. Disponível em: <http://www.brenda-enzymes.info> Acesso em: 17 dez. 2010.

BREWSTER, C.; IRIA, J.; CIRAVEGNA, F.; WILKS, Y. **The Ontology: Chimaera or Pegasus.** In: Proceedings of Dagstuhl Seminar on Machine Learning for the Semantic Web. Dagstuhl, p.13-18, 2005.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M.A. **Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis.** In: Proceedings of the 1st European Semantic Web Symposium, Greece, 2004.

BUITELAAR, P.; SINTEK, M. OntoLT Version 1.0: **Middleware for Ontology Extraction from Text.** In: Proceedings of Demo Session - International Semantic Web Conference, Hiroshima, Japan, 2004.

CAMPOS, L.M. (a). **Metodologia para Definição de Domínio no Reúso de Ontologias Biomédicas.** Qualificação de Doutorado (Doutorado em Ciência da Informação), Universidade Federal Fluminense, 175 f. 2009.

CAMPOS, L.M.; CARVALHO, M.G.P; CAMPOS, M.L.A; BRAGANHOLO, V.; CAMPOS, M.L.M. **Exploring GO Term Definitions to Enhance Automatic Annotation of Molecular Function.** In: Proceedings of International Workshop on Genomic Databases (IWGD), 2010, Buzios, Rio de Janeiro, p.1-2, 2010.

CAMPOS, M.L.A. **Em Busca de Princípios Comuns na Área de Representação da Informação: Uma Comparação Entre o Método de Classificação Facetada, o Método de Tesauro Baseado em Conceito e a Teoria Geral da Terminologia.** 198 f. Dissertação de Mestrado (Mestrado em Ciência da Informação), Universidade Federal do Rio de Janeiro, 1994.

CAMPOS, M.L.A. **Integração de Ontologias: o domínio da bioinformática.** RECIIS – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v.1, p.117-121, 2007.

CAMPOS, M.L.A. (b). **Aspectos semânticos da compatibilização terminológica entre ontologias no campo da Bioinformática.** In: X Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), João Pessoa, 2009.

CAMPOS, M.L.A. **O papel das definições na pesquisa em ontologia.** Perspectivas em Ciência da Informação, Belo Horizonte, v.15, p.10-20, 2010.

CAMPOS, M.L.A.; CAMPOS, M.L.M.; DAVILA, A.M.R.; GOMES, H.E.; CAMPOS, L.M.; LIRA, L. **Aspectos Metodológicos no Reúso de Ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos**. RECIIS – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v.3, p.64-75, 2009.

CHANDRASEKARAN, B.; JOSEPHSON, J.R.; BENJAMINS, V.R. **What Are Ontologies, and Why Do We Need Them?** IEEE Intelligent Systems, v.14, n.1 p.20-26, 1999.

CHEMICAL ENTITIES OF BIOLOGICAL INTEREST (ChEBI). Disponível em: <http://www.ebi.ac.uk/chebi> Acesso em: 17 dez. 2010.

CIMIANO, P. **Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.** New York: Springer, 375p., 2006.

CORMODE, G.; MUTHUKRISHNAN, S. **The String Edit Distance Matching Problem with Moves.** In: Proceedings of the 13th annual ACM-SIAM Symposium on Discrete algorithms, Philadelphia, USA, p.667-676, 2002.

COSTA, L.C. **Extração (Automática) de Conclusões Estruturadas de Textos de Artigos Científicos em Ciências Biomédicas.** Qualificação de Doutorado (Doutorado em Ciência da Informação). 145 f. Universidade Federal Fluminense, 2009.

COPI, I.M.; COHEN, C. **Essentials of Logic.** Upper Saddle River, Prentice Hall:Pearson, 395p., 2004.

CORCHO, O.; FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. **Methodologies, Tools and Languages for Building Ontologies. Where is Their Meeting Point?** Data & Knowledge Engineering, v.46, n.1, p.41–64, 2003.

CYBERLEX. **Tropes.** Disponível em:<http://www.cyberlex.pt/produtos.html> Acesso em: 18 ago. 2010.

CYBERLEX. **Tropes® Versão 7.0 Manual de referência Copyright ACETIC** (1994-2006), 2006.

CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V. **GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.** In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002.

CUNNINGHAM, H.; MAYNARD, D.; BONTCHEVA, K.; TABLAN, V. ; ASWANI, N.; ROBERTS, I; GORRELL, G.; FUNK, A.; ROBERTS, A.; DAMLJANOVIC, D.; HEITZ, T.; GREENWOOD, M.A; SAGGION, H.; PETRAK, J. ; LI, Y; PETERS, W. *et al.* **Developing Language Processing Components with GATE Version 5 (a User Guide).** Disponível em: < http://gate.ac.uk/sale/tao/tao.pdf > Acesso em: 1 ago. 2010.

DAHLBERG, I. **A Referent-Oriented, Analytical Concept Theory for INTERCONCEPT**. In: International Classification, v.5, n.3, p.142-150, 1978.

DAHLBERG, I. **Conceptual Definitions for INTERCONCEPT.** In: International Classification, v.8, n.1, p.16-22, 1981.

DECLERCK, T. **A Set of Tools for Integrating Linguistic and Non-Linguistic Information.** In: Proceedings of Workshop on Semantic Authoring, Annotation and Knowledge Markup, Lyon, França, 2002.

DEGGAU, R.; FILETO, R. **Enriquecendo Data Warehouses Espaciais com Descrições Semânticas.** In: Workshop de Teses e Dissertações em Bancos de Dados (WTDBD), Fortaleza, 2009.

DICIONÁRIO DO AURÉLIO ONLINE. Disponível em: <http://www.dicionariodoaurelio.com> Acesso em: 20 ago. 2010.

DING, Y.; FOO, S. **Ontology Research and Development Part 1 – A Review of Ontology Generation.** Journal of Information Science, v.28, n.2, p.123-136, 2002.

EILBECK, K.; LEWIS, S.E.; MUNGALL, C.J.; YANDELL, M.; STEIN, L.; DURBIN, R.; ASHBURNER, M. **The Sequence Ontology: a Tool for the Unification of Genome Annotations.** Genome Biology, v.6, n.5, p.1-12, 2005.

EHRIG, M.; SURE, Y. **FOAM – Framework for Ontology Alignment and Mapping Results of the Ontology Alignment Evaluation Initiative.** In: Proceedings of the Conference on Knowledge Capture (K-CAP) - Workshop on Integrating Ontologies, 2005.

ELIAS, A.B. **Expansão Semântica de Consultas Baseada em Ontologias: Uma Experimentação no Domínio Biomédico.** 129 f. Dissertação (Mestrado em Informatica) - Universidade Federal do Rio de Janeiro, 2009.

FENSEL, D.; van HARMELEN, F.; HORROCKS, I.; McGUINNESS, D.L.; PATEL-SCHNEIDER, P.F. **OIL: An Ontology Infrastructure for the Semantic Web.** IEEE Intelligent Systems, v.16, n.2, p.38-45, 2001.

FLYBASE. Disponível em:<flybase.org > Acesso em: 15 dez. 2010

FREITAS, F.; SCHULZ, S.; MORAES, E. **Pesquisa de Terminologias e Ontologias Atuais em Biologia e Medicina.** RECIIS- Revista Eletrônica de Cominicação, Informação & Inovação em Saúde, v.3, n.1, p.8-20, 2009.

FREITAS, M.C. **Elaboração Automática de Ontologias de Domínio: Discussão e Resultados.** 142f. Tese de Doutorado (Doutorado em Letras). Pontifícia Universidade Católica do Rio de Janeiro, 2007.

FREITAS, M.C.; QUENTAL, V. **Subsídios para a Elaboração Automática de Taxonomias.** In: Proceedings of Workshop em Tecnologia da Informação e da Linguagem Humana, p.1585-1594, 2007.

FRISHMAN, D.; VALENCIA, A. **BIOSAPIENS: A European Network of Excellence to Develop Genome Annotation Resources.** FRISHMAN, D.; VALENCIA, A. (eds.) Modern Genome Annotation: The Biosapiens Network, New York: Springer, 490p., 2008.

FUNDEL, K.; KUFFNER, R.; ZIMMER, R. **RelEx—Relation Extraction Using Dependency Parse Trees.** Bioinformatics, v.23, n.3, p.365-371, 2007.

GAVA, T.B.S.; MENEZES, C.S. **Uma Ontologia de Domínio para a Aprendizagem Cooperativa.** In: XIV Simpósio Brasileiro de Informática na Educação, Rio de Janeiro, p.1-12, 2003.

GATE. Disponível em: < http://gate.ac.uk > Acesso em: 19 set. 2010.

GENE ONTOLOGY (a) disponível em: <http://www.geneontology.org> Acesso em: 1 mai. 2010.

GENE ONTOLOGY (b). An Introduction to the Gene Ontology Disponível em: <http://www.geneontology.org/GO.doc.shtml> Acesso em: 1 fev. 2010.

GENE ONTOLOGY (c). AmiGO. Disponível em: <http://amigo.geneontology.org/cgi-bin/amigo/go.cgi> Acesso em: 24 fev. 2010.

GENE ONTOLOGY (d). Wiki OBOL Disponível em: <http://wiki.geneontology.org/index.php/Obol> Acesso em: 27 dez 2010.

GENE ONTOLOGY (e). Extended GO Ontology Relations. Disponível em: <http://www.geneontology.org/GO.ontology-ext.relations.shtml> Acesso em: 7 jul. 2010.

GENE ONTOLOGY CONSORTIUM. **The Gene Ontology (GO) Database and Informatics Resource.** Nucleic Acids Research, v.32, p.258-261, 2004.

GLOCK, H. **Does Ontology Exist?** Philosophy, v.77, p.235-260, 2002.

GOMES, H.; CAMPOS, M.L.A. **Contribuição da Teoria da Terminologia na Elaboração de Hiperdocumentos com Função de Tesauro.** In: IV Simpósio Iberoamericano de Terminología (RITerm), Buenos Aires, Argentina, 2004.

GÓMEZ-PÉREZ, A.; BENJAMINS, V.R. **Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods.** In: Proceedings of the IJCAI Workshop on Ontologies and Problem-Solving Methods, p.1-15, 1999.

GOODMAN, N. **Biological Data Becomes Computer Literate: New Advances in Bioinformatics.** Current Opinion in Biotechnology, v.13, n.1, p.68-71, 2002.

GRUBER,T.R. **A Translation Approach to Portable Ontology Specifications.** Knowledge Acquisition, v.5, n.2, p.199-220, 1993.

GRUBER, T. **Ontology.** LING, L.L. ÖZSU,M.T. (eds.) Entry in the Encyclopedia of Database Systems. New York:Springer-Verlag, 2008.

GU, H.; WEI, D.; MEJINO-JR, J.L.V.; ELHANAN, G. **Relationship Auditing of The FMA Ontology.** Journal of Biomedical Informatics, v. 42, n.3, p.550-557, 2009.

GUARINO, N. **Formal Ontology, Conceptual Analysis and Knowledge Representation.** International Journal of Human and Computer Studies, v.43, n.5-6, p.625-640, 1995.

GUARINO, N. **Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration.** In: Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information (SCIE), p.139-170, 1997.

GUARINO, N. (a). **Formal Ontology and Information Systems.** In: Proceedings of Formal Ontology in Information Systems (FOIS'98), Trento, Italy, p. 3-15, 1998.

GUARINO, N. (b). **Some Ontological Principles for Designing Upper Level Lexical Resources.** In: Proceedings of 1st International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

GUARINO, N.; GIARETTA, P. **Ontologies and Knowledge Bases Towards a Terminological Clarification.** In: Towards Very Large Knowledge Bases - Knowledge Building and Knowledge Sharing, Amsterdam, p. 25-32, 1995.

GUIMARÃES, M.P.; CAVALCANTI, M.C.R. **Proveniência de dados na área de Bioinformática.** Relátorio Técnico (Monografias em Sistemas e Computação), ISSN 1982-9035, 2009.

HAAV, H.; LUBI, T.A. **Survey of Concept-Based Information Retrieval Tools on The Web.** In: Proceedings of 5th East-European Conference ADBIS, 2001.

HAKENBERG, J.; LEAMAN, R.; VO, N.H.; JONNALAGADDA, S.; SULLIVAN, R.; MILLER, C.; TARI, L.; BARAL, C.; GONZALEZ, G. **Efficient Extraction of Protein-Protein Interactions from Full-Text Articles.** IEEE-ACM Transactions on Computational Biology and Bioinformatics, v.7, n.3, p.481-494, 2010.

HEARST, M.A. **Automatic Acquisition of Hyponyms from Large Text Corpora.** In: Proceedings of the 14th conference on Computational Linguistics, Stroudsburg, PA, USA, v.2, p.539-545, 1992.

HEARST, M.A. **Automated Discovery of WordNet Relations.** In: WordNet: An Electronic Lexical Database and Some of its Applications, p.1-26, 1998.

HJORLAND, B.; NICOLAISEN, J. **The Epistemological Lifeboat Epistemology and Philosophy of Science for Information Scientists.** Disponível em: <http://www.db.dk/jni/lifeboat> Acesso em: 11 jan. 2010.

HOGAN, W.R. **What's in an 'is a' link?** In: Proceedings of International Conference on Biomedical Ontology, Buffalo, 2009.

HWANG, C.H. **Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information.** In: Proceedings of Knowledge Representation Meets Databases, 1999.

IMMUNOLOGY DATABASE AND ANALYSIS PORTAL. Disponível em: <www.immport.org> Acesso em: 10 dez. 2010.

INTEGRATING NETWORK OBJECTS WITH HIERARCHIES (INOH). Disponível em: <http://www.inoh.org> Acesso em: 25 abr. 2010.

KRALLINGER, M.; VALENCIA, A.; HIRSCHMAN, L. **Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology.** Genome Biology, v. 9, n.8, p.1-14, 2008.

KÖHLER, J.; MUNN, K.; RUEGG, A.; SKUSA, A.; SMITH, B. **Quality Control for Terms and Definitions in Ontologies and Taxonomies.** BMC Bioinformatics, v.7, n.212, p.1-12, 2006.

LACROIX, Z.; CRITCHLOW, T. **Bioinformatics Managing Scientific Data.** San Francisco:Morgan Kaufmann Publishers, 441p., 2003.

LEMOS, M. **Workflow para Bioinformática.** 239f. Tese de Doutorado (Doutorado em Informática). Pontifícia Universidade Católica do Rio de Janeiro, 2004.

LEMOS, M.; SEIBEL, L.F.B; CASANOVA, M.A. **BioNotes: A System for Biosequence Annotation.** In: Proceedings of 14th International Workshop on Database and Expert Systems Applications, p.1-5, 2003.

LEWIS, S.E. **Gene Ontology: Looking Backwards and Forwards.** Genome Biology, v.6, n.1, p.1-4, 2004.

LEWIS, S.E.; SEARLE, S.M.J.; HARRIS, N.; GIBSON, M.; IYER, V.; RICHTER, J.; WIEL, C.; BAYRAKTAROGLU, L.; BIRNEY, E.; CROSBY, M.A.; KAMINKER, J.S.; MATTHEWS, B.B.; PROCHNIK, S.E.; SMITH, C.D.; TUPY, J.L.; RUBIN, G.M.; MISRA, S.; MUNGALL, C.J.; CLAMP, M.E. **Apollo: a Sequence Annotation Editor.** Genome Biology, v.3, n.12, p.1-14, 2002.

MAEDCHE, A.; STAAB, S. **Measuring Similarity Between Ontologies.** In: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, Ontologies and the Semantic Web (EKAW), 251-263, 2002.

MATHIAK,B.; ECKSTEIN,S. **Five Steps To Text Mining In Biomedical Literature.** In: Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics, 2004.

METAMAP. Disponível em: <http://metamap.nlm.nih.gov> Acesso em: 22 dez. 2010.

METAMAP TRANSFER (MMTx) Disponível em: <http://mmtx.nlm.nih.gov> Acesso em: 22 dez. 2010.

MICHAEL, J.; MEJINO-JR., J.L.V; ROSSE, C. **The Role of Definitions in Biomedical Concept Representation.** In: Proceedings of AMIA Symposium, p.463-467, 2001.

MORIN, E.; JACQUEMIN, C. **Automatic Acquisition and Expansion of Hypernym Links.** Computer and the Humanities, 2003.

MILLER, G.A. **WordNet: A Lexical Database for English.** Communications of the ACM, v.38, n.11, p.39-41, 1995.

MOUSE GENOME INFORMATICS. Disponível em: <http://www.informatics.jax.org> Acesso em: 15 dez. 2010.

MUNGALL ,C.J. **Obol: Integrating Language and Meaning in Bio-Ontologies.** Comparative and Functional Genomics, v.5, n.6-7, p.509-520, 2004.

MUNGALL, C.J.; BADA, M.; BERARDINI, T.Z.; DEEGAN, J.; IRELAND, A.; HARRIS, M.A.; HILL, D.P.; LOMAX, J. **Cross-product Extensions of the Gene Ontology.** Journal of Biomedical Informatics, 2010.

NATURAL LANGUAGE TOOLKIT (NLTK). Disponível em: <http://www.nltk.org> Acesso em: 2 fev. 2010.

NARDI, J.C.; FALBO, R.A. **Uma Ontologia de Requisitos de Software.** In: Proceedings of 9º Workshop Iberoamericano de Ingeniería de Requisitos y Ambientes de Software, p.111-124, 2006.

NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY. Disponível em: <bioontology.org> Acesso em: 2 jun. 2010.

NOY ,N.F. **Semantic Integration: A Survey Of Ontology-Based Approaches.** SIGMOD Record. v.33, n.4, p. 65-70, 2004.

NOY, N.F.; McGUINNESS, D.L. **Ontology Development 101: A Guide to Creating Your First Ontology.** Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

NOY, N.F.; MUSEN, M.A. **PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment.** In: Proceedings of the 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. p.450-455, 2000.

OBO - Open Biomedical Ontologies (a). The Open Biological and Biomedical Ontologies Disponível em: <www.obofoundry.org> Acesso em: 29 jul. 2010.

OBO - Open Biomedical Ontologies (b). OBO Foundry Principles. Disponível em: <http://www.obofoundry.org/crit.shtml> Acesso em: 29 jul. 2010.

OBO - Open Biomedical Ontologies (c). OBO Relation Ontology Disponível em: <http://www.obofoundry.org/ro/#OBO_REL> Acesso em: 12 set. 2010 .

OGREN, P.V.; COHEN, K.B.; ACQUAAH-MENSAH, G.K.; EBERLEIN, J.; HUNTER, L. **The Compositional Structure of Gene Ontology Terms.** In: Proceedings of the Pacific Symposium on Biocomputing, v.9, p.214-225, 2004.

OGREN, P.V.; COHEN, K.B.; HUNTER, L. **Implications of Compositionality in the Gene Ontology for Its Curation and Usage.** In: Proceedings of the Pacific Symposium on Biocomputing, p.174-185, 2005.

ONTOLOGY RESEARCH GROUP. Disponível em: <org.buffalo.edu> Acesso em: 23 set. 2010.

PALAKAL, M.; MUKHOPADHYAY, S.; STEPHENS, M. **Identification of Biological Relationships from Text Documents.** In: CHEN, H.; FULLER, S.S.; FRIEDMAN, C.; HERSHP, W.(eds.) Medical Informatics Knowledge Management and Data Mining in Biomedicine. New York: Springer-Verlag, USA, p.449-489, 2005.

PATHWAY ONTOLOGY Disponível em: <http://rgd.mcw.edu/tools/ontology/ont_search.cgi> Acesso em: 17 dez. 2010.

PEREIRA, B.O.; LÔBO, R.M. Detectando Padrões e Estruturas nas Descrições Textuais dos Termos da Gene Ontology. 98 f. Projeto Final (Graduação em Ciência da Computação), Universidade Federal do Rio de Janeiro, 2010.

PICKLER, M.E.V. **Web Semântica: Ontologias como Ferramentas de Representação do Conhecimento.** Perspectiva em Ciência da Informação, Belo Horizonte, v.12, n.1, p. 65-83, 2007.

PREFUSE. Disponível em: < http://prefuse.org> Acesso em: 27 out. 2010.

PUBMED. Disponível em: <www.ncbi.nlm.nih.gov/pubmed> Acesso em: 18 nov. 2010.

RINALDI, F.; SCHNEIDER, G.; KALJURAND, K.; DOWDALL, J.; ANDRONIS, C.; PERSIDIS, A.; KONSTANTI, O. **Mining Relations in the GENIA Corpus**. In: Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, 2004.

RINALDI, F.; SCHNEIDER, G.; KALJURAND, K.; HESS, M.; ANDRONIS, C.; KONSTANDI, O.; PERSIDIS, A. **Mining of Relations Between Proteins over Biomedical Scientific Literature Using a Deep-Linguistic Approach.** Journal Artificial Intelligence in Medicine, v.39, n.2, p.127-136, 2007.

SACCHAROMYCES GENOME DATABASE. Disponível em: <www.yeastgenome.org> Acesso em: 10 dez. 2010.

SALES, L.F. **Ontologias de Domínio Estudo das Relações Conceituais e sua Aplicação.** 140f. Dissertação (Mestrado em Ciência da Informação). Universidade Federal Fluminense, 2006.

SALES, L.F. **Relações Conceituais para Instrumentos de Padronização Terminológica: um novo modelo para o uso em Ontologias.** In: VIII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), Salvador, Bahia, 2007.

SALES, L.F.; CAMPOS, M.L.A; GOMES, H.E. **Ontologias de Domínio: Um Estudo das Relações Conceituais.** Perspectivas em Ciência da Informação, Belo Horizonte, v.13, n.2, 2008.

SCHULZ, S.; KUMAR, A.; BITTNER, T. **Biomedical Ontologies: What part-of is and isn't.** Journal of Biomedical Informatics, v.39, n. 3, p. 350-361, 2006.

SAS INSTITUTE INC.SAS Text Miner. Disponível em: <http://www.sas.com/text-analytics/text-miner> Acesso em: 29 dez. 2010.

SINTEK, M.; BUITELAAR, P.; OLEJNIK, D. **A Formalization of Ontology Learning From Text.** In: Proceedings of the Workshop on Evaluation of Ontology-based Tools - International Semantic Web Conference, 2004.

SILVA, V.S. **Uma Abordagem para Alinhamento de Ontologias Biomédicas para Apoiar a Anotação Genômica.** 110f. Dissertação (Mestrado em Informatica), Universidade Federal do Rio de Janeiro, 2010.

SMITH, B. **Ontology and Information Systems.** Disponível em: <http://ontology.buffalo.edu/ontology_long.pdf> Acesso em: 1 mai.2010.

SMITH, B.; ASHBURNER, M.; ROSSE, C.; BARD, J.; BUG, W.; CEUSTERS, W.; GOLDBERG, L.J.; EILBECK, K.; IRELAND, A.; MUNGALL, C.J; THE OBI CONSORTIUM; LEONTIS, N.; ROCCA-SERRA, P.; RUTTENBERG, A.; SANSONE, S.; SCHEUERMANN, R.H.; SHAH, N.; WHETZEL, P.L.; LEWIS, S. **The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration**. Nature Biotechnology, v.25, p.1251-1255, 2007.

SMITH, B.; CEUSTERS, W.; KLAGGES, B.; KÖHLER, J.; KUMAR, A.; LOMAX, J.; MUNGALL, C.; NEUHAUS, F.; RECTOR, A.L.; ROSSE, C. **Relations in Biomedical Ontologies.** Genome Biology, v.6, n.5, p.1-15, 2005

SMITH, B.; KÖHLER, J.; KUMAR, A. **On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology.** In: Database Integration in the Life Sciences, p.1-17, 2004.

SMITH, B.; KUMAR, A. **On Controlled Vocabularies in Bioinformatics: A Case Study in the Gene Ontology.** In: Drug Discovery Today - BIOSILICO, v.2, n.6, p.246-252, 2004.

SMITH, B.; ROSSE, C. **The Role of Foundational Relations in the Alignment of Biomedical Ontologies.** In: Proceedings of World Congress on Medical Informatics (MEDINFO), v.11, n.1 p.444-448, 2004.

SMITH, B.; WILLIAMS, J.; SCHULZE-KREMER, S. **The Ontology of the Gene Ontology.** In: Proceedings of AMIA Symposium, p.609-613, 2003.

SPECIALIST NLP Tools. Disponível em : <http://lexsrv3.nlm.nih.gov/Specialist> Acesso em: 18 dez. 2010.

STANFORD. **Protégé.** Disponível em: <protege.stanford.edu> Acesso em: 4 mai. 2010.

TARCZY-HORNOCH, P.; MINIE, M. **Bioinformatics Challenges and Opportunities.** In: CHEN, H.; FULLER, S.S.; FRIEDMAN, C.; HERSHP, W. (eds.) Medical Informatics Knowledge Management and Data Mining in Biomedicine. New York:Springer-Verlag, USA, p.63-94, 2005.

THE ARABIDOPSIS INFORMATION RESOURCE (TAIR). Disponível em: <www.arabidopsis.org> Acesso em: 20 out. 2010.

UNIFIED MEDICAL LANGUAGE SYSTEM (UMLS). Disponível em: <http://www.nlm.nih.gov/research/umls> Acesso em: 18 dez. 2010.

UMLS METATHESAURUS. Disponível em: <http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html> Acesso em: 18 dez. 2010.

USCHOLD, M. **Building Ontologies: Towards a Unified Methodology.** In: Proceedings of 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, p.1-20, Cambridge, UK,1996.

USCHOLD, M.; GRUNINGER, M. **Ontologies: Principles, Methods and Applications.** Knowledge Engineering Review, v.11, n.2, p.1-69, 1996.

van HEIJST, G.; SCHREIBER, A.; WIELINGA, B.J. **Using Explicit Ontologies in KBS Development.** International Journal of Human-Computer Studies. v.46, n.2-3, p.183-292, 1997.

WAGNER, G. **Geração e Análise Comparativa de Sequências Genômicas de Trypanosoma Rangeli.** 105f. Dissertação de Mestrado (Mestrado em Biologia Celular e Molecular). Fundação Oswaldo Cruz, 2006.

WELTY, C.; GUARINO, N. **Supporting Ontological Analysis of Taxonomic Relationships.** Data & Knowledge Engineering, v.39, n.1, p.51-74, 2001.

WROE, C. J.; STEVENS, R.; GOBLE, C.A.; ASHBURNER, M. **Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL.** In: Pacific Symposium on Biocomputing, p.624-635, 2003.

ANEXO A – EXEMPLOS DE DEFINIÇÕES PADRONIZADAS DA *Gene Ontology*

Fonte: (Gene Ontology, 2010b)

Termo: [x] envelope
Definição Padrão: The double lipid bilayer enclosing the [x] and separating its contents from the rest of the cytoplasm.

Termo: [x] membrane, única membrana.
Definição Padrão: The lipid bilayer surrounding a(n) [x].

Termo: [x] membrane, membrana dupla.
Definição Padrão: Either of the lipid bilayers that enclose the [x] and form the [x] envelope.

Termo: [x] inner membrane
Definição Padrão: The inner, i.e. lumen-facing, lipid bilayer of the [x] envelope.

Termo: [x] outer membrane
Definição Padrão: The outer, i.e. cytoplasm-facing, lipid bilayer of the [x] envelope.

Termo: [x] membrane lumen
Definição Padrão: The region between the inner and outer lipid bilayers of the [x] envelope.

Termo: [x] development
Definição Padrão: The process whose specific outcome is the progression of the [x] over time, from its formation to the mature structure.

Termo: embryonic [x] morphogenesis
Definição Padrão: The process, occurring in the embryo, by which the anatomical structures of [x] are generated and organized.

Termo: larval [x] morphogenesis

Definição Padrão: The process, occurring in the larva, by which the anatomical structures of [x] are generated and organized.

Termo: post-embryonic [x] morphogenesis
Definição Padrão: The process, occurring after embryonic development [and prior to (e.g. larval) development], by which the anatomical structures of [x] are generated and organized.

Termo: [x] formation
Definição Padrão: The process that gives rise to [x]. This process pertains to the initial formation of a structure from unspecified parts.

Termo: [x] structural organization
Definição Padrão: The process that contributes to creating the structural organization of [x]. This process pertains to the physical shaping of a rudimentary structure.

Termo: [x] maturation
Definição Padrão: A developmental process, independent of morphogenetic (shape) change, that is required for [x] to attain its fully functional state. [descrição de x].

Termo: [x] cell differentiation
Definição Padrão: The process whereby a relatively unspecialized cell acquires specialized features of a [x] cell.

Termo: [x] cell fate commitment
Definição Padrão: The process whereby the developmental fate of a cell becomes restricted such that it will develop into a [x] cell.

Termo: [x] cell fate specification
Definição Padrão: The process whereby a cell becomes capable of differentiating autonomously into a [x] cell in an environment that is neutral with respect to the developmental pathway.

Termo: [x] cell fate determination

Definição Padrão: The process whereby a cell becomes capable of differentiating autonomously into a [x] cell regardless of its environment

Termo: [x] cell development
Definição Padrão: The process aimed at the progression of a [x] cell over time, from initial commitment of the cell to a specific fate, to the fully functional differentiated cell.

Termo: [x] cell maturation
Definição Padrão: A developmental process, independent of morphogenetic (shape) change, that is required for a [x] cell to attain its fully functional state. [descrição de x]

Termo: [x] other organism during symbiotic interaction
Definição Padrão: [descrição de x], where the two organisms are in a symbiotic relationship.

Termo: [x] host
Definição Padrão: [descrição de x]. The host is defined as the larger of the organisms involved in a symbiotic interaction.

Termo: [x] symbiont
Definição Padrão: [descrição de x]. The symbiont is defined as the smaller of the organisms involved in a symbiotic interaction.

Termo: [x] metabolic process
Definição Padrão: The chemical reactions and pathways involving [x], [descrição de x].

Termo: [x] biosynthetic process
Definição Padrão: The chemical reactions and pathways resulting in the formation of [x]. [descrição de x].

Termo: [x] catabolic process
Definição Padrão: The chemical reactions and pathways resulting in the breakdown of [x], [descrição de x].

Termo: [x] fermentation

Definição Padrão: The enzymatic conversion of [x] to simpler components, resulting in energy in the form of adenosine triphosphate (ATP).

Termo: [x] salvage
Definição Padrão: Any process that produces [x] from derivatives of it, without de novo synthesis.

Termo: 'de novo' [x] biosynthetic process
Definição Padrão: The chemical reactions and pathways resulting in the formation of [x] from simpler precursors.

Termo: regulation of [processo]
Definição Padrão: Any process that modulates the frequency, rate or extent of [processo], [definição do processo].

Termo: regulation of [enzima] activity
Definição Padrão: Any process that modulates the frequency, rate or extent of [enzima] activity, the catalysis of the reaction: [reação].

Termo: regulation of [x]
Definição Padrão: Any process that modulates the frequency, rate or extent of [x], [definição de x].

Termo: regulation of [x]
Definição Padrão: Any process that modulates [x], [descrição de x].

Termo: negative regulation of [processo]
Definição Padrão: Any process that stops, prevents or reduces the frequency, rate or extent of [processo], [definição do processo].

Termo: positive regulation of [processo]
Definição Padrão: Any process that activates, maintains or increases the frequency, rate or extent of [processo], [definição do processo].

Termo: activation of [processo]
Definição Padrão: Any process that starts the inactive process [processo], [definição do processo].

Termo: activation of [enzima] activity
Definição Padrão: Any process that initiates the activity of the inactive enzyme [enzima]. The initiation of the activity of the inactive enzyme [enzima] by [processo].

Termo: maintenance of [processo]
Definição Padrão: Any process that maintains the frequency, rate or extent of the active process [processo], [definição do processo].

Termo: maintenance of [x]
Definição Padrão: Any process that maintains [x], [descrição de x].

Termo: upregulation of [processo]
Definição Padrão: Any process that increases the frequency, rate or extent of the active process [processo], [definição do process].

Termo: downregulation of [processo]
Definição Padrão: Any process that reduces the frequency, rate or extent of, but does not terminate, the active process [processo], [definição do processo].

Termo: inhibition of [processo]
Definição Padrão: Any process that prevents the activation of the inactive process [processo], [definição do processo].

Termo: termination of [processo]
Definição Padrão: Any process that stops the active process [processo], [definição do processo].

Termo: inactivation of [enzima] activity
Definição Padrão: Any process that terminates the activity of the active enzyme [enzima]. The termination of the activity of the active enzyme [enzima] by [processo].

Termo: response to [x] stimulus
Definição Padrão: A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a [x] stimulus.

Termo: cellular response to [x] stimulus
Definição Padrão: A change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a [x] stimulus.

Termo: behavioral response to [x]
Definição Padrão: A change in the behavior of an organism as a result of a [x].

Termo: age-dependent response to [x] stimulus
Definição Padrão: A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a [x] stimulus, where the change varies according to the age of the cell or organism.

Termo: detection of [x] stimulus
Definição Padrão: The series of events in which a [x] stimulus is received and converted into a molecular signal.

Termo: (família) sensory perception
Definição Padrão: The series of events required for an organism to receive a sensory stimulus, convert it to a molecular signal, and recognize and characterize the signal.

Termo: sensory perception of [x] stimulus
Definição Padrão: The series of events required for an organism to receive a [x] stimulus, convert it to a molecular signal, and recognize and characterize the signal.

Termo: detection of [x] stimulus during sensory perception of [y]
Definição Padrão: The series of events during the perception of [y] in which a [x] stimulus is received and converted to a molecular signal.

Termo: downstream [efeito conhecido], [iniciador desconhecido]
Definição Padrão: The series of molecular signals that result in [...]

Termo: events in [caminho conhecido], [iniciador e alvo final desconhecidos]
Definição Padrão: The series of molecular signals involving [...]

Termo: [x] signaling pathway
Padrão: The series of molecular signals generated as a consequence of [x] binding to [descrição]

Termo: [x] localization
Definição Padrão: The processes by which [x] (where [x] is a substance or cellular entity, such as a protein complex or organelle) is transported to, and/or maintained in, a specific location.

Termo: establishment of [x] localization
Definição Padrão: The directed movement of [x] to a specific location.

Termo: maintenance of [x] localization
Definição Padrão: The processes by which [x] is maintained in a location and prevented from moving elsewhere.

Termo: [x] secretion
Definição Padrão: The regulated release of [x] from a cell or group of cells.

Termo: [x] export
Definição Padrão: The directed movement of [x] out of a cell or organelle.

Termo: [x] import
Definição Padrão: The directed movement of [x] into a cell or organelle.

Termo: establishment of [x] orientation
Definição Padrão: The processes that set the alignment of [x] relative to other cellular structures.

Termo: (família) L-amino acid
Definição Padrão: L-Y is the levorotatory isomer of [x].

Termo: (família) D-amino acid
Definição Padrão: D-Y is the dextrorotatory isomer of [x].

Termo: (família) constitutive activity
Definição Padrão: This activity is constitutive and therfore always present, regardless of demand.

Termo: (família) inducible activity
Definição Padrão: This activity is inducible and therefore only present when there is demand.

Termo: (família) high affinity
Definição Padrão: In high affinity transport the transporter is able to bind the solute even if it is only present at very low concentrations.

Termo: (família) low affinity
Definição Padrão: In low affinity transport the transporter is able to bind the solute only if it is present at very high concentrations.

Termo: [x] transmembrane transporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other. [x] is [descrição de x].

Termo: [x] uptake transmembrane transporter activity
Definição Padrão: Catalysis of the transfer of [x] from the outside of a cell to the inside of a cell across a membrane. [x] is [descrição de x].

Termo: [x] efflux transmembrane transporter activity
Definição Padrão: Catalysis of the transfer of [x] from the inside of a cell to the outside of the cell across a membrane. [x] is [descrição de x].

Termo: active transmembrane [x] transporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, up the solute's concentration gradient. The transporter binds the solute and undergoes a

series of conformational changes. Transport works equally well in either direction. [x] is [descrição de x].

Termo: primary active transmembrane [x] transporter activity
Definição Padrão: Catalysis of the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by a primary energy source. Primary energy sources known to be coupled to transport are chemical, electrical and solar sources. [x] is [descrição de x].

Termo: decarboxylation-driven active transmembrane [x] transporter activity
Definição Padrão: Enables the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by decarboxylation of a cytoplasmic substrate. [x] is [descrição de x].

Termo: light-driven active transmembrane [x] transporter activity
Definição Padrão: Enables the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by light. [x] is [descrição de x].

Termo: methyl transfer-driven active transmembrane [x] transporter activity
Definição Padrão: Enables the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by a methyl transfer reaction. [x] is [descrição de x].

Termo: oxidoreduction-driven active transmembrane [x] transporter activity
Definição Padrão: Enables the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by an exothermic flow of electrons from a reduced substrate to an oxidized substrate. [x] is [descrição de x].

Termo: P-P-bond-hydrolysis-driven transmembrane [x] transporter activity
Definição Padrão: Enables the transport of [x] across a membrane, up the solute's concentration gradient, by binding the solute and undergoing a series of conformational changes. Transport works equally well in either direction and is driven by the hydrolysis of the diphosphate bond of inorganic pyrophosphate, ATP, or another nucleoside triphosphate. The transport protein may or may not be transiently phosphorylated, but the substrate is not phosphorylated. [x] is [descrição de x].

Termo: secondary active transmembrane [x] transporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, up the solute's concentration gradient. The transporter binds the solute and undergoes a series of conformational changes. Transport works equally well in either direction and is driven by a chemiosmotic source of energy. Chemiosmotic sources of energy include uniport, symport or antiport. [x] is [descrição de x].

Termo: [x]: solute antiporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, up the solute's concentration gradient. The transporter binds the solute and undergoes a series of conformational changes. Transport works equally well in either direction and is driven by a antiport mechanism whereby two or more species are transported in opposite directions in a tightly coupled process not directly linked to a form of energy other than chemiosmotic energy. [x] is [descrição de x].

Termo: [x]: solute symporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, up the solute's concentration gradient. The transporter binds the solute and undergoes a series of conformational changes. Transport works equally well in either direction and is driven by a symport mechanism whereby two or more species are transported together in the same direction in a tightly coupled process not directly linked to a form of energy other than chemiosmotic energy. [x] is [descrição de x].

Termo: [x] uniporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, up the solute's concentration gradient. The transporter binds the solute and undergoes a

series of conformational changes. Transport works equally well in either direction and is driven by a uniport mechanism which is independent of the movement of any other molecular species. [x] is [descrição de x].

Termo: passive transmembrane [x] transporter activity
Definição Padrão: Catalysis of the transfer of [x] from one side of the membrane to the other, down the solute's concentration gradient. [x] is [descrição de x].

Termo: [x] channel activity
Definição Padrão: Catalysis of facilitated diffusion of [x] (by an energy-independent process) by passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.

Termo: gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens in response to a specific stimulus.

Termo: dephosphorylation-gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens in response to dephosphorylation of one of its constituent parts.

Termo: ion-gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens in response to a specific ion stimulus.

Termo: ligand-gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens when a specific ligand has been bound by the channel complex or one of its constituent parts.

Termo: mechanically-gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens in response to a mechanical stress.

Termo: phosphorylation-gated [x] channel activity

Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel that opens in response to phosphorylation of one of its constituent parts.

Termo: voltage-gated [x] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [x] by a channel whose open state is dependent on the voltage across the membrane in which it is embedded.

Termo: [x] channel activity
Definição Padrão: Catalysis of facilitated diffusion of [x] (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.

Termo: [x] gated [y] channel activity
Definição Padrão: Catalysis of the transmembrane transfer of [y] by a channel that opens in response to stimulus by [x]. Transport by a channel involves catalysis of facilitated diffusion of a solute (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel, without evidence for a carrier-mediated mechanism.

Termo: Qualquer fase do ciclo celular
Definição Padrão: Progression through [nome da fase], [descrição da fase].

Termo: Qualquer processo de ciclo celular
Definição Padrão: The cell cycle process whereby [descrição do processo].

Termo: (família) membrane fusion
Definição Padrão: The joining of the lipid bilayer membrane around [x] to the lipid bilayer membrane around [y].

Termo: [x] assembly
Definição Padrão: The aggregation, arrangement and bonding together of a set of components to form a [x].

Termo: [x] distribution
Definição Padrão: Any process that establishes the spatial arrangement of [x].

Termo: [x] binding

Definição Padrão: Interacting selectively and non-covalently with [x], [descrição de x].

Termo: [enzima] activity

Definição Padrão: Catalysis of the reaction: [reação].

Termo: [x] receptor activity

Definição Padrão: Combining with [x] to initiate a change in cell activity.

Termo: [x] transporter activity

Definição Padrão: Enables the directed movement of [x] into, out of or within a cell, or between cells.

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

CCMN - Bloco C - Cidade Universitária - Ilha do Fundão
Rio de Janeiro - RJ CEP: 21941-916
www.ppgi.ufrj.br