IX CONGRESS CONGRESO
ISKO-SPAIN ISKO-ESPAÑA
Valencia 11th, 12th, 13th March 2009 11, 12 y 13 de Marzo de 2009
New Perspectives for the organisation and dissemination of knowledge
Nuevas perspectivas para la difusión y organización del conocimiento

INTEGRAÇÃO DE ONTOLOGIAS EM DOMÍNIO INTERDISCIPLINAR: EXPERIÊNCIA NO CAMPO DA BIOMEDICINA

Maria Luiza de Almeida Campos[1], Maria Luiza Machado Campos[2], Linair Maria Campos[3]

1 UFF-GCI-PPGI-UFF – Rua Lara Vilela, 126, Niterói – RJ - Brasil - marialuizalmeida@gmail.com,
2 UFRJ-PPGI-IM/NCE - Av. Athos da Silveira Ramos, 274, Rio de Janeiro, RJ, Brasil – mluiza@ufrj.br,
3 (PPGCI-UFF/IBICT) – Rua Lara Vilela, 126, Niterói – RJ - Brasil - linair@nce.ufrj.br

RESUMO:

Ontologias têm tido seu uso difundido para a integração de informações e recursos, em especial na área de Biomedicina, caracterizada por sua temática complexa e pela disponibilidade de diversas ontologias com grande número de termos e mudanças constantes. O objetivo desse trabalho é discutir uma proposta metodológica para a integração dessas ontologias com o foco em seu reuso, abrangendo desde a análise do domínio até o seu uso articulado, com o apoio de ferramentas de software. A proposta é conduzida no contexto de um experimento na área de Biomedicina. Seus resultados preliminares apontam para a atualidade dos aportes teóricos da Ciência da Informação, aliados à Ciência da Computação, e a pertinência do papel do cientista da informação no cenário da pesquisa em organização da informação apoiada por ontologias.

ABSTRACT:

Ontologies have been widely used to integrate information and resources, especially in Biomedicine, where ontologies are diverse, complex, big, and change frequently. The goal of this work is a methodological approach to integrate such ontologies focusing on their reuse, covering aspects from domain analysis through their articulated use, supported by software tools. Our proposal is conducted in the context of an experiment in the Biomedicine area and its preliminary results point to the actuality of Information Science theoretical contributions, allied with Computer Science, and the relevance of the information scientist role in the research scenario of information organization supported by ontologies.

PALAVRAS CHAVES:

Ontologias, integração, mapeamento, biomedicina

1 . INTRODUÇÃO

Nos últimos anos, no campo da genômica, iniciativas da comunidade científica internacional levaram a um crescimento explosivo de informações biológicas geradas todos os dias. A preocupação inicial, então, foi a criação e manutenção de bancos de dados para armazenar essas informações biológicas. Rapidamente os bancos de dados genômicos tiveram seu uso ampliado, com muitos genomas seqüenciados. O foco das pesquisas começou a se transferir do mapeamento dos genomas para a análise da vasta gama de informações resultantes da caracterização funcional dos genes através da Biologia Molecular e da Bioinformática, esta última uma área interdisciplinar na qual Biologia, Ciência da Computação e Tecnologia da Informação fazem parte. Tornou-se fundamental a interligação entre os dados obtidos pelos diversos projetos de pesquisa ao redor do mundo, envolvendo o inter-relacionamento de enzimas, genes, componentes químicos, doenças, espécies, tipos de células, órgãos, etc.

Como as fontes de informação científica crescem rapidamente, o profissional da informação tem papel fundamental na organização e recuperação da informação científica, no levantamento de fontes de informação de dados para o estudo dos organismos de interesse, no levantamento de estratégias para integração desses dados aos já existentes no laboratório e na harmonização de conceitos para tratamento e intercâmbio de informações (Heidorn, 2007). Assim, para que equipes e/ou instituições troquem recursos científicos entre si é preciso encontrar uma forma comum de descrição e acesso a estes recursos, de modo a facilitar a busca e integração dos mesmos. Neste sentido, as *ontologias* ganham importância fundamental para garantir a harmonização semântica e a recuperação de informações.

Como apresentado em diversos estudos, ontologias (Gruber, 1993; Guarino, 1998; Vickery, 1997; Smith, 2002) surgem no âmbito da Inteligência Artificial, na década de 90, como instrumento de representação de conhecimento. Quando o conhecimento de um domínio é representado em uma linguagem declarativa, o conjunto de objetos que podem ser representados é chamado de universo do discurso. Foi nesse sentido que surgiram as ontologias, com o intuito de descrever dados manipulados por programas, através da definição de um conjunto de termos que pudessem representar domínios e tarefas a serem executadas por estes programas. Uma ontologia é, assim, um conjunto de conceitos padronizados, termos e definições aceitas por uma comunidade particular. A mais freqüente definição de ontologia é a de Gruber (1993), para o qual "uma ontologia é uma especificação de uma conceituação". Desta forma, uma ontologia diferentemente de seu significado na área de Filosofia, é considerada um artefato tecnológico.

Uma importante iniciativa de construção de ontologias para Biomedicina é a OBO (Open Biomedical Ontologies) (OBO, 2005), um consórcio que disponibiliza vocabulários controlados para uso compartilhado em diferentes domínios ligados à

medicina e à biologia. Um dos vocabulários mais utilizados existentes hoje na OBO é a Gene Ontology (GO) (Gene Ontology Consortium, 2001).

A GO possibilita descrever aspectos referentes a processos biológicos, funções moleculares e componentes celulares relacionados aos genes de organismos de diferentes espécies. A GO, entretanto, não fornece uma visão completa dos descritores relacionados a outros aspectos do genoma, nem das estruturas das seqüências de proteínas que são expressas pelos genes. Nesse sentido, ontologias de temáticas tais como a de tecidos, vias metabólicas e dados de alinhamentos de seqüências genômicas são úteis para complementar a anotação e o estudo dos genomas e vêm sendo produzidas pela comunidade biomédica em resposta à demanda por descritores que forneçam uma visão mais rica e precisa dos experimentos da área.

Essas ontologias, entretanto, apresentam problemas em relação à forma como estão estruturadas e à forma como foram estabelecidos os recortes de seus domínios. Apesar do consórcio da OBO estar trabalhando em uma reformulação de um conjunto de ontologias da área, a tarefa é complexa e demanda tempo. Sendo assim, a fim de se obter em curto prazo um conjunto de descritores complementares à GO, é preciso lidar com os problemas encontrados nesses vocabulários, dentre eles a ocorrência de um mesmo termo definido e/ou organizado de forma distinta em diferentes ontologias da área. Essa ocorrência deve ser encarada de três formas: (i) o termo que consta da GO está corretamente enquadrado em seu domínio; (ii) o termo é mais adequado no domínio da outra ontologia onde foi encontrado; (iii) o termo não é adequado a nenhum dos dois domínios.

Além disso, ao considerarmos a compatibilização de ontologias de temáticas afins, outros problemas de natureza semântica se apresentam entre essas ontologias, como, por exemplo: o uso de diferentes terminologias para denominar o mesmo conceito; um mesmo conceito sendo expresso por um único termo em uma ontologia e por mais de um termo em outra ontologia; a existência de diferentes níveis de detalhamento na formação de hierarquias onde os conceitos semelhantes se inserem; ou, ainda, o uso de diferentes relações para o mesmo conceito nas diferentes ontologias.

Ding e Foo (2002) apresentaram um levantamento de alguns trabalhos sobre integração de ontologias, tanto manuais como semi-automáticos, os quais manipulam os conceitos e também suas relações. Estes trabalhos utilizam várias estratégias computacionais, que vão desde o uso de algoritmos de clusterização para identificar relações candidatas, até uma variedade de técnicas com o objetivo de estabelecer a proximidade de dois conceitos, baseadas na estrutura hierárquica da ontologia. Todas essas estratégias sofrem de algum modo da falta de semântica fornecida pelas estratégias lingüísticas, e poderiam se beneficiar de mecanismos para melhorar a caracterização dos conceitos. Essas estratégias, ainda, têm seu foco voltado para o aumento da eficiência do tratamento computacional da integração, carecendo de uma

visão mais abrangente que contextualize o domínio de conhecimento e o usuário, assim como suas necessidades de informação e seus objetivos ao utilizar as ontologias.

Este artigo pretende apresentar, em torno da problemática da integração de ontologias, uma atividade de pesquisa que vem sendo realizada pelo grupo de pesquisa "Ontologia e taxonomia: aspectos teóricos e metodológicos", na qual ações interdisciplinares envolvendo conhecimentos da Ciência da Informação e da Ciência da Computação têm possibilitado desenvolvimento de pesquisas conjuntas. Nosso experimento se dá no âmbito do consórcio BioWebDB (www.biowebdb.org), envolvendo o Instituto Oswaldo Cruz e as Universidades Federais do Rio de Janeiro e Santa Catarina. Este consórcio, inicialmente financiado pelo CNPq, reúne um grupo de pesquisadores na área de Biologia, Bioinformática, Computação e Ciência da Informação em torno dos estudos de genômica comparativa (que compreende a análise e comparação de genomas de diferentes espécies) e bancos de dados genômicos. Nosso espaço de atuação no grupo é identificar e propor uma solução para a integração de um conjunto de ontologias de interesse, além da Gene Ontology, que serão utilizadas para a anotação de genomas de tripanosomatídeos. Nesse contexto, nossa participação tem atingido alguns resultados preliminares que apresentaremos a seguir.

2. INTEGRAÇÃO DE ONTOLOGIAS

Ontologias podem ser reutilizadas de diversas formas, que ora resultam na criação de uma ontologia independente a partir dos conceitos de outras (podendo ser estendidos e adaptados), ora preservam as ontologias originais. Na literatura, essas formas de reuso têm sido contempladas por abordagens que recebem diferentes denominações e interpretações. Dentre estas, as mais comuns são: alinhamento, junção, mapeamento e integração. Estas abordagens, embora conduzam a resultados distintos, possuem aspectos comuns, como, por exemplo, a preocupação em encontrar termos semelhantes nas ontologias sendo consideradas.

Assim, o alinhamento difere da junção em relação ao seu resultado: em vez de gerar uma ontologia adicional, resultado da combinação das ontologias reutilizadas, o alinhamento mantém as ontologias reutilizadas inalteradas e em seus locais de origem, gerando um conjunto de vínculos (links) entre essas ontologias. Esses vínculos contêm um conjunto de informações sobre como compatibilizar as ontologias reutilizadas e são expressos em um modelo persistente (que existe fisicamente) em separado.

O conjunto de vínculos expressos em um modelo persistente produzido pelo processo de alinhamento é um *mapeamento* (mapping) entre as ontologias. As informações contidas no mapeamento vão depender do tipo de vínculo semântico encontrado entre os elementos e do tipo de formalismo utilizado na ontologia para representar a sua semântica. Por exemplo, dois elementos podem ser semelhantes (em

diferentes graus): ou um pode ser parte do outro, ou então podem ter algum outro tipo de relacionamento que é identificado com o auxílio de um especialista no domínio.

Mapeamentos de semelhança podem expressar diferentes graus de similaridade. (Felicíssimo; Breitman, 2004) (Kalfoglou; Schorlemmer, 2003) (Su, 2004). Para se determinar o grau de similaridade, geralmente diversos fatores são levados em conta, tais como: similaridade lingüística entre os termos, compatibilidade dos seus atributos, posicionamento do termo na estrutura hierárquica da ontologia, dentre outros (Bruijn; Ehrig; Feier 2006). Um dos aspectos do mapeamento é a questão de como achar os candidatos. Este problema é chamado de *obtenção de correspondências* ou *casamento* (matching).

A obtenção de correspondências pode basear-se em diferentes tipos de técnicas para estimar os candidatos: (i) **lingüístico**: baseado na semelhança dos nomes dos termos e relações, e em suas definições textuais e sinônimos incluídos nas próprias ontologias; (ii) **estrutural**: baseado na estrutura da ontologia, como, por exemplo, levando-se em conta o posicionamento dos termos na estrutura hierárquica das ontologias sendo comparadas ou então as suas relações partitivas ou ainda outros tipos de relações que sejam utilizadas de forma semelhante nas ontologias comparadas (Euzenat; Shvaiko, 2007); (iii) **contextual**: baseado na adição de conhecimento suplementar, como, por exemplo, o derivado de recursos descritos com as ontologias, se houver, ou nas informações de uma terceira ontologia ou vocabulário que possua uma hierarquia de conceitos, como a Wordnet (Miller, 1990), que possa ser utilizada, por exemplo, para procura de sinônimos ou apoio para aferir similaridade semântica entre conceitos (Cross e Wang, 2005) (Sabou; D'aquin; Motta, 2006).

Por fim, no contexto desse trabalho, definimos integração como uma aplicação onde se consideram ontologias de mesmo domínio alinhadas formando um modelo persistente (contido em um mapeamento), o qual serve como ponto de partida para interligar, quando possível, as diferentes ontologias consideradas.

Com base na revisão de literatura das áreas da Ciência da Informação e da Ciência da Computação, identificamos que existem três aspectos do reuso que devem ser cobertos em uma proposta de integração de ontologias: (i) o estudo do domínio, das fontes e necessidades de informação relacionadas aos interesses do usuário final, onde se inserem as ontologias sendo reutilizadas; (ii) os aspectos operacionais que dizem respeito ao estabelecimento de correspondências entre as ontologias; (iii) o apoio computacional ao reuso das ontologias de forma integrada, para o usuário final, voltado para a tomada de decisão no emprego dessas ontologias e não apenas no apoio a tarefas de cunho operacional referentes à compatibilização de vocabulários.

3. TRABALHOS RELACIONADOS À INTEGRAÇÃO DE ONTOLOGIAS

Os trabalhos relacionados à integração de ontologias têm sido objeto de estudos da Ciência da Computação há algum tempo. Em nossas atividades interdisciplinares percebemos que apesar do *lócus* da integração de ontologias estar no âmbito da Ciência da Computação, estudos na área da Ciência da Informação relacionados à compatibilização de linguagens, já na década de 80 do século passado, apresentavam princípios que consideramos fundamentais e que podem ser aplicados no contexto das ontologias. O presente estudo pretende resgatar conteúdos teóricos e metodológicos desta área que, agregados a conteúdos da Ciência da Computação, permitirão o desenvolvimento de aplicativos de integração onde aspectos semânticos serão ressaltados.

3.1 CONTRIBUIÇÕES DA CIÊNCIA DA INFORMAÇÃO

No domínio da Ciência da Informação, estudos de natureza metodológica para apoiar o levantamento de termos que compõem as unidades de um dado domínio de conhecimento, têm sido objeto de pesquisa de muitos estudiosos (Soergel, 1982; Dahlberg, 1978; Hjorland, 2002). Além destes, destacamos a proposta de Shatford (1986) para a descrição de fontes de informação imagéticas. Esses estudos forneceram diretrizes sistemáticas que têm sido investigadas, no contexto deste trabalho, para uma análise preliminar do domínio.

Além da análise do domínio, a Ciência da Informação tem estudado o problema da compatibilização de vocabulários no âmbito das linguagens de indexação de conteúdos, como os tesauros. Nesse cenário inserem-se os trabalhos de Dahlberg (1983), a qual propõe a construção de uma matriz de compatibilidade conceitual, através de seu método analítico-sintético. A matriz de compatibilidade conceitual é um mapeamento da potencialidade semântica das linguagens estudadas, fornecendo os resultados da análise de compatibilidade entre linguagens sob os pontos de vista semântico e estrutural. A compatibilidade entre linguagens, segundo Dalhberg, compreende três fases: (i) a coincidência conceitual – quando dois conceitos combinam suas características – grau de equivalência; (ii) correspondência conceitual - dois conceitos combinam a maior parte de suas características – similaridade; (iii) correlação conceitual - dois conceitos são correlacionados através de símbolos matemáticos, estabelecendo uma medida de correlação, quando possuem diferentes níveis de detalhe, ou quando a relação entre eles não é de semelhança.

Outros trabalhos publicados na área da Ciência da Informação que versam sobre o tema de compatibilização são: o método de reconciliação de tesauros proposto por Neville (1972), e a ampliação de escopo de tesauros proposta por Rada e Martin (1987).

O método de Neville baseia-se no princípio que se devem compatibilizar os conceitos (os conteúdos conceituais dos descritores, que estão expressos pelas

definições) e não os descritores somente. Esse método propõe uma abordagem de linguagem intermediária, baseada na codificação numérica de conceitos, através da qual se torna possível o estabelecimento da equivalência conceitual de descritores de diferentes linguagens, considerando ainda o tratamento de correspondência de um para *n* entre os termos a compatibilizar.

Rada e Martin (1987) relatam um projeto na área biomédica que detecta propriedades herdadas de um tesauro (armazenado em um banco de dados) e usa essas propriedades para guiar especialistas na tomada de decisão sobre como ampliar o escopo de tesauros de temáticas afins, através da sua interligação, para indexar literatura médica. Os autores propõem ainda a herança de atributos comuns de um tesauro para outro, como é o caso de relacionamentos entre termos, e consideram o uso de sinônimos na identificação de termos semelhantes.

Shatford, por sua vez, propõe um modelo para a descrição de imagens que captura os objetos e ações representadas por estas, e ainda seus elementos. Esses aspectos são descritos, de modo geral e detalhado, através de um quadro de referência, o qual possui facetas para capturar a descrição da imagem em relação aos objetos que ela representa, seus aspectos temporais e espaciais, dentre outros. Esse modelo de referência, devidamente adaptado, pode ser aplicado como requisito para a análise de domínio.

Observamos ainda que, embora muitos dos trabalhos da área de Ciência da Informação datem da década de 1980, ainda se mostram bastante pertinentes e atuais, como demonstram o amplo detalhamento das questões de compatibilização um para *n*, de Neville, a proposta do registro de conceito de Dahlberg e o modelo de Shatford.

3.2 CONTRIBUIÇÕES DA CIÊNCIA DA COMPUTAÇÃO

A literatura sobre reuso de ontologias na Ciência da Computação, na qual se insere a integração, explora com detalhes os diferentes aspectos envolvidos do ponto de vista operacional, ou seja, do *que* necessita ser feito ou tratado, e os problemas que são enfrentados nesse contexto. Em relação aos aspectos metodológicos, sobre *como* fazer o reuso, o que se encontra com mais freqüência diz respeito aos aspectos computacionais, como, por exemplo, os algoritmos mais eficazes para promover a compatibilidade entre ontologias, tanto em relação à precisão de seus resultados como em relação à sua rapidez (Noy, Musen, 2000).

Em relação aos algoritmos para compatibilidade, destacam-se os voltados para encontrar termos correspondentes entre ontologias, sendo que essa correspondência pode representar uma relação de semelhança ou outra qualquer, que possa ser obtida, por exemplo, a partir da análise do contexto de uso da ontologia. Diversas propostas que implementam algoritmos para busca de correspondências são encontradas na

literatura (Euzenat; Shvaiko, 2007). Estas possuem aspectos que são resumidos na Tabela 1, conforme a perspectiva apresentada por Ehrig e colegas (2004).

Os algoritmos que tratam de compatibilidade muitas vezes são implementados através de recursos de software (ferramentas ou ambientes de software) voltados para o reuso de ontologias. Esses recursos possuem meios para interagir com o usuário de modo que este possa ajustar as sugestões feitas pelo software para a junção ou mapeamento de termos.

Tabela 1. – aspectos relacionados à busca por correspondência entre ontologias

Tipo da camada	Comparação baseada em	Resumo das abordagens adotadas para aferir similaridade
Dados	Valores de dados tais como nomes ou números inteiros. Foco em aspectos lingüísticos.	**Similaridade sintática**: para nomes, determinam-se quantas ações são necessárias para transformar um nome no outro.
Ontologia	Relações semânticas entre os conceitos, regras, restrições, axiomas, instâncias. Foco em aspectos estruturais.	**Similaridade taxonômica**: usa a estrutura hierárquica da ontologia - conceitos em níveis mais altos são mais genéricos e semanticamente menos semelhantes que os conceitos em níveis mais baixos; **Similaridade de relações**: compara as classes de origem (domain) e de destino (range) - relações com mesma origem e destino tendem a ser iguais; **Similaridade de conjuntos de conceitos**: em vez de calcular semelhança entre dois conceitos, calcula entre conjuntos de conceitos, agrupados em clusters.
Contexto	Uso externo dos elementos da ontologia; assume-se que elementos semelhantes possuem padrões semelhantes de uso. Foco no contexto.	A aferição da similaridade vai depender do contexto. Por exemplo, se estamos avaliando similaridade entre livros, estes podem ser considerados semelhantes se seus autores forem os mesmos. A similaridade é calculada com base nos *atributos* dos elementos da ontologia sendo comparados e não nos elementos em si.

Esses softwares costumam abordar aspectos complementares da correspondência entre ontologias e, dentro de cada aspecto, nem sempre exploram todas as suas possibilidades. Essa questão é ilustrada por Su (2004) em uma pesquisa do estado da arte de ferramentas e ambientes para mapeamento de ontologias. Por exemplo, dentre um subconjunto dos softwares que utilizam aspectos lingüísticos e estruturais, alguns levam em conta para o mapeamento apenas o termo (FCA-Merge e Glue), enquanto outros levam em conta o termo e suas relações (Chimaera e Momis), havendo ainda os que utilizam o termo, suas relações e restrições (Prompt). Além disso, alguns softwares, como o Chimaera, utilizam aspectos lingüísticos e estruturais, mas não exploram os aspectos contextuais. Em relação aos resultados desses softwares, alguns, como o Prompt, geram uma nova ontologia, fruto de uma junção, enquanto que outros, como o Momis e o Glue, geram uma lista de pares de termos semelhantes.

Entretanto, nenhum desses softwares está inserido em propostas que coloquem o reuso em uma perspectiva mais abrangente do que os aspectos operacionais de se encontrar termos semelhantes e permitir ao usuário fundi-los ou mapeá-los.

Nesse sentido, alguns autores chegam a propor tarefas mais gerais que são necessárias no processo de reuso. Gangemi, Steve e Giancomelli (1996), por exemplo, afirmam que é necessário identificar os termos básicos e suas definições necessárias e suficientes em forma textual, porém não sugerem como fazer essa identificação, nem quais princípios adotar para construir as definições. Pinto e Martins (2001), por sua vez, sugerem que o reuso começa na seleção de ontologias a serem reutilizadas. Os autores fornecem uma série de diretrizes importantes sobre como escolher as ontologias a serem reutilizadas, como, por exemplo, se estas possuem documentação e se possuem nível de detalhe adequado. Sugerem ainda uma série de tarefas a serem feitas a partir e sobre a seleção das ontologias desejadas: que as definições de seus conceitos devem ser melhoradas, as relações maximizadas, e as ontologias avaliadas por especialistas, tanto do ponto de vista técnico quanto do ponto de vista do seu uso. E ainda, que o conhecimento essencial deve ser identificado. Entretanto, não fornecem muitos detalhes sobre como devem se dar essas tarefas.

4. ETAPAS DA INTEGRAÇÃO DE ONTOLOGIAS: UMA PROPOSTA METODOLÓGICA NO CAMPO DA BIOMEDICINA

Os nossos estudos têm apontado para a importância da investigação no âmbito dos estudos em Compatibilização de Linguagens, na Ciência da Informação e na Ciência da Computação como apontado anteriormente. Consideramos que a partir deles podemos obter diretrizes teóricas e metodológicas para o reuso em ontologias (Campos, 2004). Com base nessa perspectiva e nos três aspectos de reuso identificados na seção 3.1, estruturamos nossa proposta de integração de ontologias, conforme a Figura 2. Por motivos de espaço, neste artigo vamos privilegiar o detalhamento de alguns aspectos de nossa proposta, para os quais temos protótipo de software em desenvolvimento, em detrimento de outros.

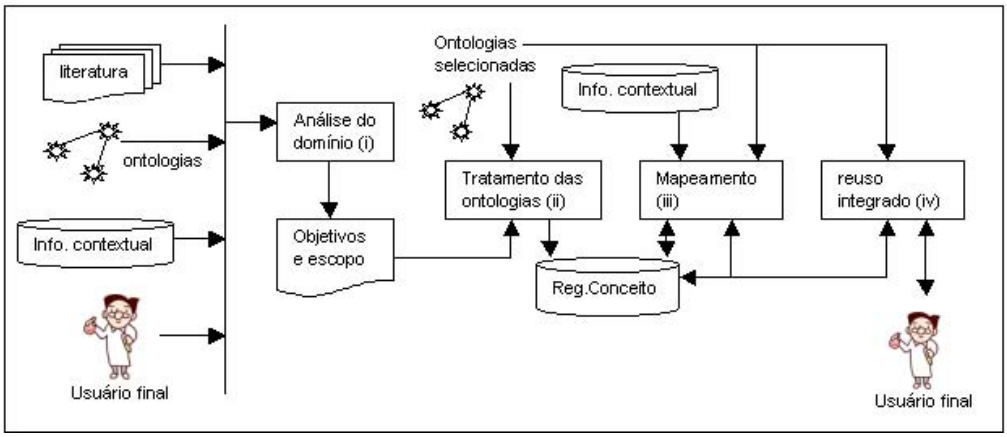

Figura 2. - Etapas da integração de ontologias

No âmbito deste trabalho, o mapeamento das ontologias está estreitamente ligado ao reuso integrado que é feito pelo usuário final dentro da comunidade local. É a partir da experiência do reuso local que o usuário se insere de forma colaborativa na comunidade global. Nesse contexto, a análise do domínio fornece ao cientista da informação uma visão geral dos objetivos desse usuário, de modo a que se possa selecionar um conjunto de ontologias de interesse e, dentro destas, o escopo dos descritores a serem considerados e integrados. Dessa forma, espera-se fornecer uma solução abrangente para o reuso integrado das ontologias de interesse, como ilustrado na Figura 2 e detalhado a seguir.

4.1. ANÁLISE DO DOMÍNIO

A análise do domínio de conhecimento aonde a ação informacional irá se debruçar é de fundamental importância para a definição do *corpus terminológico* que se irá determinar para as ações de integração. Latour (1997), na teoria ator-rede, estabelece que a ciência deva ser estudada na prática dos cientistas, incluindo a relação homem – máquina e sociedade. A ciência se faz nas bancadas dos laboratórios, definindo no processo da ação o seu conteúdo e todo o contexto em que esses atores atuam no social. Nesse sentido, é fundamental que tenhamos a visão do domínio de interesse a partir da nossa participação ativa dentro dele. Para isso, uma das abordagens que adotamos é a participação em uma série de seminários e entrevistas, que ajudam a compreender melhor o domínio alvo, no caso o estudo genômico no âmbito da Biologia Molecular de Tripanosomatídeos e Flebotomíneos no Laboratório de Biologia Molecular de Tripanosomatídeos e Flebotomíneos do Instituto Oswaldo Cruz (IOC). Partindo dessa capacitação inicial, onde se dá uma familiarização com o ambiente do usuário final, deve-se proceder ao levantamento das fontes de informação.

Hjorland (2002) apresenta que em Ciência da Informação existem recursos informacionais que devem ser identificados, descritos, organizados e comunicados para atender a objetivos específicos e que ela pode se beneficiar ao considerar a visão analítica do domínio, por meio de abordagens diversas, tais como: análise de literatura especializada, levantamento de ferramentas computacionais, dentre outros. Em nossa proposta, a análise da literatura especializada é feita através da análise do curriculum Lattes dos pesquisadores seniores do Laboratório. Essa análise aponta os temas de interesse, as linhas de pesquisa e os principais artigos. A compreensão desse material fornece uma base para a seleção preliminar de ontologias de interesse, que devem ser validadas junto ao usuário final. Para isso, essas ontologias devem ter a sua utilidade evidenciada.

Para identificar a utilidade de uma ontologia é importante que os interessados possam entender sua finalidade, os princípios metodológicos que conduziram à sua construção, e ainda o contexto em que esta foi desenvolvida. Para isso, no nosso entender, é preciso explicitar uma série de critérios utilizados na construção da ontologia, os quais envolvem aspectos tais como: (i) a identificação dos vocabulários

ou fontes de informação usadas para o domínio sendo representado; (ii) os propósitos a serem atingidos, os quais determinam o seu uso; (iii) a determinação do nível de detalhe (grão) da ontologia; (iv) o seu grau de formalismo, dentre outros.

Além disso, é importante dar uma visão de alto nível das classes das ontologias, uma vez que estas muitas vezes contêm centenas e às vezes milhares de termos de grande profundidade (dez ou mais níveis hierárquicos são comuns). Por exemplo, o termo "des-arginine_complement_5a" na ontologia MESH está no nono nível hierárquico, sendo que um dos termos que está no terceiro nível hierárquico possui 83 irmãos, cada um com vários níveis de profundidade.

As perspectivas utilizadas para a construção de uma ontologia podem estar implícitas ou, ao contrário, explicitadas através de algum formalismo ou descrição livre. Caso estejam explícitas, no nosso entender trará benefícios para a sua compreensão dentro da análise do domínio, na medida em que fornece elementos claros e precisos sobre o seu conteúdo. A análise do conteúdo e da estrutura da ontologia provê assim os insumos para o seu tratamento mais preciso em um sistema informatizado e auxiliando a tomada de decisão do pesquisador quanto a sua possível utilidade.

O problema que se apresenta nesse contexto é como descrever as ontologias de interesse e como determinar seus aspectos relevantes. Para responder a esse questionamento, propomos o modelo de Shatford (1986). Este, uma vez adaptado, pode fornecer uma base adequada para a descrição desses atributos, uma vez que provê uma série de facetas sob as quais pode-se olhar um objeto de análise. Shatford desmembra os seus níveis em categorias, inspirada pela teoria da Classificação Facetada de Ranganathan (1967). As categorias são: QUEM, ONDE, QUANDO e COMO/O QUE. Nesse sentido, situamos nossa adaptação do modelo de Shatford para a descrição de ontologias com base na analogia que é possível fazer com os três níveis de detalhes de Shatford. Cabe aqui um alerta ao traçar nosso paralelo com as imagens. Em Shatford, estamos descrevendo uma imagem de cada vez; em nossa proposta, estamos descrevendo uma ontologia (e não uma classe da ontologia) de cada vez.

Em relação ao "DE Genérico", entendemos que este se aplica perfeitamente ao propósito da descrição do conteúdo de uma ontologia, na medida em que, de maneira análoga a uma imagem, nela também estão contidos objetos e ações, que podem ser identificados de modo genérico como sendo os grandes temas, ou ramos que fazem parte da ontologia.

Em relação ao "DE Específico", uma adaptação se faz necessária. Em Shatford esse nível é utilizado para descrever um aspecto específico da imagem. Entretanto, como nosso foco não é identificar o conteúdo específico de uma classe e sim da ontologia como um todo, o nível "DE Específico" deve retratar um aspecto que nos permita individualizar de alguma maneira este todo, de modo a atingir nosso objetivo de caracterizar bem a ontologia. Para isso, a descrição das classes de nível mais alto da

ontologia, bem como das relações presentes na mesma, cumpre este propósito, na medida em que fornece detalhes que permitem ao interessado reconhecer sua utilidade.

O nível "SOBRE" possui o mesmo intuito que o modelo original, ou seja, captar a interpretação do conteúdo da ontologia, levando em conta o seu contexto.

Além disso, apesar de as categorias de Shatford fornecerem elementos para a determinação do contexto em que uma imagem foi criada, julgamos importante criar duas categorias adicionais para caracterizar de forma mais precisa as ontologias. Para isso, desmembramos a categoria COMO/O QUE em duas categorias separadas, COMO e O QUE, acrescentando ainda a categoria PARA QUE. Além disso, a coluna SOBRE é utilizada para apresentar detalhes adicionais, muitas vezes em forma de ponteiros para documentos extensos (como, por exemplo, a documentação sobre a metodologia utilizada), que podem ser úteis para compreender melhor a ontologia.

Tabela 2. - explicação sobre o conteúdo da ontologia, adaptado do modelo de Shatford

Categoria	DE Genérico	DE Específico	SOBRE
QUEM	ramos principais da ontologia	classes de primeiro nível dentro de cada ramo principal e as relações presentes na ontologia.	link onde se podem encontrar informações mais detalhadas sobre a ontologia.
ONDE	instituição responsável pela ontologia.	detalhes do contexto (ex: um projeto de pesquisa), onde se insere a ontologia.	link onde se podem encontrar informações mais detalhadas sobre instituição responsável.
QUANDO	ano última versão	dia e mês da última versão.	código da última versão.
O QUE	tema da ontologia	detalhes do tema da ontologia.	agrupamento temático da ontologia
COMO	tipo do formalismo e metodologia.	documentação da Metodologia utilizada.	fontes utilizadas para a definição / estruturação dos termos da ontologia
PARA QUE	propósito da ontologia	de que maneira tem sido usada e os benefícios que a ontologia traz ao ser usada.	bancos genômicos que as utilizam e informações adicionais obtidas a partir de anotações com a ontologia.

A categoria COMO é utilizada para que se descreva como a ontologia foi desenvolvida, ou seja, qual a metodologia e as fontes de informação utilizadas (possivelmente outras ontologias, ou literatura especializada). A categoria PARA QUE é utilizada para que se descreva o propósito para o qual a ontologia foi desenvolvida.

A partir da análise da literatura especializada podemos selecionar um subconjunto de ontologias de interesse, juntamente com uma série de informações contextuais (e.g.. artigos, bancos de dados que utilizam as ontologias) que são validadas junto ao usuário final. Após essa validação, temos os insumos para estabelecer o escopo e os objetivos que irão nortear o tratamento das ontologias.

4.2. TRATAMENTO DAS ONTOLOGIAS

O registro do conceito é o ponto central da integração das ontologias. Ele fornece uma visão geral dos conceitos relevantes de um domínio, como um dicionário específico de um domínio, permitindo apontar quais vocabulários se referem a cada conceito e com que termos.

Devido à grande complexidade e ao grande número de conceitos de determinados domínios, o registro de conceito deve ser materializado a partir de

ontologias já existentes e complementado pelo usuário final aos poucos, na medida de sua necessidade, quando as ontologias existentes não possuírem os descritores desejados, devendo ser apoiado por ferramenta de software.

É no registro do conceito que ficam registrados os dados sobre a natureza dos termos contidos nas ontologias reutilizadas. Essa natureza é dada a partir de uma ontologia de alto nível, que no nosso caso é a Basic Formal Ontology (BFO) (Grenon, Smith e Goldberg, 2004), a qual é voltada para representar as estruturas básicas da realidade. Ela possui duas perspectivas básicas para agrupar as entidades que existem no mundo: (i) **continuantes (continuant)**: aquelas que perduram com o tempo, embora venham a sofrer mudanças, como, por exemplo: os rins, os cromossomos, as células; (ii) **ocorrentes (occurrent)**: aquelas que acontecem no mundo, sendo referidas comumente como processos, eventos, atividades, como, por exemplo: a síntese de proteína, a difusão de uma epidemia (Grenon, Smith e Goldberg, 2004).

Essa ontologia foi escolhida em razão de estar sendo adotada na OBO, dentro do contexto de reformulação das ontologias desse consórcio, utilizadas em nossos experimentos. Além da ontologia de alto nível, deve-se utilizar uma ontologia de relações para o domínio, de modo a facilitar o estabelecimento de novas relações na etapa de mapeamento. Essa ontologia de relações é escolhida ou montada na etapa de análise do domínio.

O registro do conceito é carregado inicialmente com um subconjunto escolhido dos termos das ontologias reutilizadas. O escopo de tal escolha se dá como resultado da análise do domínio. Após a etapa de carga, o usuário informa para cada ontologia a vinculação à BFO, a qual fica gravada no registro de conceito e não em cada ontologia, pois as ontologias mudam com freqüência, o que implicaria em redundância de esforços a cada nova versão. Através do registro de conceito é possível atualizar apenas os termos que sofreram mudanças (detectadas através de heurísticas cujo detalhamento foge ao escopo deste trabalho), minimizando esforços.

A vinculação à BFO é feita a partir do tratamento prévio dos ramos ou das classes de primeiro nível das ontologias sendo reutilizadas, através da associação dessas classes a uma propriedade que vamos denominar **semantic_role** (Byrne e McCracken, 1999), a qual deve ser incluída, caso não haja nenhuma semelhante, na ontologia de relações do domínio. As subclasses dos ramos ou classes de primeiro nível herdam a natureza da vinculação da classe mãe. Naturalmente, algum grau de imprecisão pode ocorrer, nos casos em que a hierarquia da ontologia estiver mal formada em relação à natureza da relação **is-a** (gênero/espécie). Entretanto, nossos experimentos evidenciam que o ganho de precisão obtido é maior do que se a ontologia não fosse tratada previamente. Cabe ressaltar que, devido ao grande número de termos das ontologias envolvidas, a vinculação manual de cada termo individualmente em detrimento do tratamento apenas dos termos de mais alto nível é inviável, por consumir muito tempo e

estar mais sujeita a erros pelo grande escopo da tarefa. A vinculação à BFO é apoiada por ferramentas de software, que atualizam o registro de conceito.

4.3. MAPEAMENTO

Um dos aspectos da integração é a *compatibilidade* entre os vocabulários reutilizados, na qual o mapeamento ocupa um papel central. Neste sentido, é importante deixar bem claro que o uso que ora fazemos do termo tem seu campo definido no âmbito da Ciência da Informação e é um estudo seminal desta área, envolvendo teóricos como Soergel (1982), Dalhberg (1981), Neville (1972) e Glushkov e colegas (1978). Para estes últimos, compatibilidade é a medida de similaridade entre duas linguagens, na qual se introduz o conceito de graus de compatibilidade e se estabelece a distinção entre compatibilidade em plano semântico e no plano lingüístico.

Em nossa proposta, estendemos o trabalho de Dahlberg (1981, 1983) para aferir o grau de compatibilidade entre os termos de forma mais precisa. Além disso, o uso de recursos computacionais também favorece o estabelecimento de casamentos de forma mais rápida, viabilizando o trabalho em grandes vocabulários. Isso é feito na etapa de mapeamento.

O mapeamento pode envolver diferentes estratégias para estimar o grau de compatibilidade entre os termos de ontologias afins e permitir ao usuário final uma compreensão mais precisa das ontologias reutilizadas, revelando: (i) a natureza dos eventuais conflitos que possam existir, cujas naturezas citamos na seção 1; e (ii) possíveis relacionamentos não explicitados entre as ontologias reutilizadas. Essas estratégias envolvem três passos consecutivos, a saber: identificação de casamentos entre termos, tratamento de conflitos e descoberta de novos relacionamentos.

Na identificação de casamentos nosso objetivo é encontrar os termos que denotam o mesmo conceito ou conceitos semelhantes e estabelecer um registro da correspondência entre esses termos nas diferentes ontologias onde ocorrem. Esse registro é gravado no registro de conceitos.

Nossa estratégia para a busca de casamentos parte da comparação dos termos pela sua semelhança lingüística e também utiliza conhecimento sobre a natureza do termo, de acordo com a sua vinculação a uma ontologia de alto nível, conforme explicado na etapa de tratamento das ontologias, item 4,.2. Cabe ressaltar que o casamento é estabelecido do conceito para os termos (um para *n*) e não entre cada uma das ontologias (*n* para *n*).

No tratamento de conflitos, nosso objetivo é gravar no registro de conceitos dados para que o usuário possa complementar, no momento do reuso integrado, item 4.4., detalhes sobre o problema ou para que possa optar por um termo como preferencial em detrimento de outro, com base nas informações fornecidas. Escolher

um termo como preferencial implica em designá-lo como representante de um conceito, no registro de conceito.

Nossa estratégia para o tratamento dos conflitos parte do registro de correspondência dos termos semelhantes e complementa suas informações com dados sobre os possíveis conflitos entre esses termos, como, por exemplo, diferenças no nível de detalhe ou na formação da estrutura hierárquica, quando for o caso.

No presente trabalho ainda estamos aprofundando o experimento da descoberta de novos relacionamentos. Para isso, explora-se a definição textual dos termos das ontologias e o conhecimento presente nos recursos descritos com essas ontologias, caso haja.

Nossa proposta de compatibilização utiliza os dados do registro do conceito para gerar uma medida numérica que expressa o grau de semelhança entre os conceitos mapeados. Em nosso aplicativo protótipo, que implementa a proposta, essa medida em um primeiro momento contempla apenas o tratamento da equivalência de um para um, embora possa haver casos em que um termo em uma ontologia venha a corresponder a mais de um termo em outra ontologia.

O registro do conceito ajuda o usuário final a ter uma visão geral de como um determinado conceito está representado em um conjunto de ontologias de interesse, de modo que ele possa compreender de forma mais precisa o contexto em que cada termo associado se insere e quais os aspectos da definição do conceito correspondente que ele privilegia.

4.3.1 ESTABELECIMENTO DE CASAMENTOS

Em nossa proposta, o casamento dos termos é concebido para ser implementado de forma desacoplada, de modo que algoritmos diversos para casamento possam ser empregados, desde que cumpram o propósito de detectar termos semelhantes e estejam de acordo com um conjunto mínimo de especificações de como atualizar o registro de conceitos, com os atributos que identificamos como relevantes. Essas especificações, cujo detalhamento não cabe aqui, explicam as interfaces de software que devem ser implementadas para atualizar o registro de conceito, o qual é armazenado em um banco de dados. Os dados ali armazenados são explorados por ferramentas de software para, dentre outras finalidades, calcular o grau de compatibilidade entre os termos sendo comparados. O registro encontra-se estruturado em uma série de tabelas dentro do banco de dados, cuja modelagem foge ao escopo deste trabalho. O registro de conceito também sofre atualização na etapa de tratamento de conflitos e de estabelecimento de novas relações. A Tabela 3 ilustra os atributos do registro de conceito.

Tabela 3. - Informações contidas no registro do conceito, estendido a partir de Dahlberg (1981)

Atributo	**Descrição**
Código único do conceito	Identificador único do conceito, utilizado para se estabelecer a correspondência deste com os termos encontrados nas ontologias a serem integradas.
Código da ontologia de origem do conceito (se houver)	Código utilizado para se referir à ontologia de origem do conceito. Se o conceito for sugerido, então a ontologia de origem não existe ou não foi identificada.
Código do termo relativo ao conceito na ontologia de origem (se houver)	Código utilizado para se referir ao termo na ontologia de origem. Se o conceito for sugerido, então o código do termo de origem não existe ou não foi identificado.
Natureza associada ao conceito	Papel semântico associado ao conceito, de acordo com a BFO.
Nome do termo relativo ao conceito	Nome do termo associado ao conceito.
Descrição textual do termo relativo ao conceito	Definição em linguagem natural, conforme a ontologia de origem ou, caso o conceito não exista em ontologias, a definição do usuário.
Fonte da definição textual	Fonte de onde foi obtida a definição do conceito.
Sinônimo(s)	Termos de conteúdo equivalente do conceito, porém que não está presente como termo na versão atual da ontologia.
Conceitos relacionados [ontologia, termo, relação]	Conceitos relacionados em outras ontologias. Para cada conceito informa-se o código da ontologia, o código do termo, o nome da relação encontrada, de acordo com uma ontologia de relações.
Código da ontologia	Código utilizado para se referir à ontologia onde é encontrado o termo que corresponde ao conceito.
Código do termo	Código utilizado para se referir ao termo que corresponde ao conceito.
Natureza associada ao termo	Natureza associada ao termo, de acordo com a BFO.
Nome do termo	Nome do termo na ontologia onde é encontrado.
Termo genérico	Termo de nível hierárquico imediatamente superior.
Termo mais alto na hierarquia	Termo de nível hierárquico mais alto dentro do ramo onde o termo se insere.
Tipo da relação encontrada entre o termo e o preferencial	Pode ser de equivalência ou então outra qualquer, que, neste caso, deve ser explicitada. Pode também não haver relação, no caso de homônimos.
Base para descoberta da relação	Lingüística, estrutural, contextual – podendo ser mais de uma destas.
Nível de detalhe (=, >, <, [código, código...])	Quando a relação é de equivalência, o termo pode ser de mesmo nível de detalhe que o conceito (=), ou então é menos (<) ou mais (>) detalhado (<), ou é necessária uma combinação do termo em questão com outro cujo código é especificado aqui, podendo haver mais de um.
Medida de similaridade	Valor aferido para a similaridade do termo em relação ao conceito.
Conflito	Tipo do conflito detectado na etapa de tratamento de conflitos.
Observação do usuário final	Comentário sobre problemas observados com o termo.

Algoritmos para casamento variam em sua complexidade. Os mais simples utilizam apenas aspectos lingüísticos para a comparação do nome dos termos. Outros, mais complexos, podem ser utilizados para obter maior precisão, como, por exemplo, o de casamento morfo-semântico e de padrões de lógica nebulosa, implementados por Rada e Martin (1987) ou ainda a abordagem para comparações entre taxonomias, de Maedche e Staab (2002). Nossa preocupação, no momento, não é em estabelecer qual o algoritmo mais preciso, mas levantar os dados que devem ser contemplados e explorados em uma solução de integração que considere as necessidades do usuário final.

4.3.2 TRATAMENTO DE CONFLITOS

De acordo com nossa abordagem, um possível conflito é detectado quando existe alguma diferença entre os termos casados. Essas diferenças podem se dar pelos seguintes motivos: (i) termos são homônimos, mas possuem natureza semântica distinta; (ii) termos são homônimos e possuem o mesmo papel semântico, mas não possuem nenhum termo genérico identificado ao longo da hierarquia que seja comum (do ponto de vista da semelhança do nome e do seu papel semântico); (iii) termos são homônimos, possuem termo genérico comum na hierarquia (do ponto de vista da semelhança do nome e do seu papel semântico), mas fazem parte de hierarquias diferentes (do ponto de vista da semelhança do nome); (iv) termos são homônimos e possuem o mesmo papel semântico, mas possuem diferentes relacionamentos associados.

O caso (i) ocorre quando conceitos diferentes são denominados por termos de mesmo nome. Como exemplo, podemos citar o casamento do conceito *excretion* (excreção) encontrado nas ontologias GO e Brenda. Na primeira, o termo possui o papel semântico de um *ocorrente* e significa "a eliminação por um organismo de dejetos que são resultado de uma atividade metabólica". Na segunda, possui o papel semântico de um *continuante*, referindo-se a um produto de uma atividade e significa "a matéria, tal como urina ou suor, que é excretada do sangue, tecidos ou órgãos".

O caso (ii) pode ocorrer por diversos motivos, dentre os quais: (a) existe um termo genérico que é comum nas duas hierarquias, mas a denominação desse termo genérico é diferente, de modo que o algoritmo de semelhança lingüística não pode detectar essa semelhança; (b) não existe o termo genérico comum porque os conceitos, embora semelhantes, são definidos de forma distinta, privilegiando aspectos distintos de suas naturezas. Por exemplo, no caso (b): na ontologia SBO um *ligand* é um tipo físico "(estrutural ou informacional) participante de um evento, ou seja, sua natureza química, ou o tipo de ação que realiza", enquanto que na GO um *ligand* é uma proteína. O caso (iii) é um caso especial do caso (ii) (b), com a diferença de que os termos genéricos comuns puderam ser identificados devido à semelhança na sua nomenclatura. O caso (iv) indica uma diferença na abordagem de definição dos termos e pode significar uma oportunidade para a migração das relações de uma ontologia para a outra, caso seja conveniente.

Outros tipos de conflito podem existir, como, por exemplo, diferentes sintaxes, mas estes não são considerados no escopo deste trabalho.

Após o tratamento de conflitos, procede-se ao cálculo do grau de compatibilidade entre termos casados de ontologias afins. Para isso, diversos valores são aferidos individualmente no registro de conceito para os diversos aspectos da comparação que é efetuada, a saber: (i) semelhança do nome; (ii) semelhança semântica em relação à estrutura hierárquica; (iii) natureza do conceito associado ao termo; (iv)

diferença no nível de detalhamento; (v) diferença em relação aos relacionamentos associados ao termo, caso haja. A fórmula para cada um desses valores depende do algoritmo que está sendo utilizado. Ao final é computado o grau de compatibilidade geral para o termo, o qual é gravado no registro de conceito.

4.3.3 ESTABELECIMENTO DE NOVAS RELAÇÕES

O estabelecimento de novas relações parte da análise das definições textuais dos termos onde é possível, com o apoio de ferramentas de mineração de texto e de uma ontologia de relações, extraindo sentenças que sugiram possíveis novos relacionamentos entre os conceitos utilizados no domínio. A proposta é que a partir da descrição do conceito possamos identificar enunciados onde possamos evidenciar novos termos. Esta conduta metodológica se apóia na Teoria do Conceito de Dahlberg (1978b). Segundo a Teoria do Conceito, as características relevantes do conceito são os elementos constitutivos da definição. O ponto principal no estabelecimento das definições dos conceitos está, portanto, na identificação das características. Ela fornece um padrão para definição, classificando-as conforme a categoria do conceito (definição genérica, partitiva, funcional) (Dahlberg, 1983). A definição genérica permite identificar a categoria do conceito, a partitiva os componentes do conceito definido, e a funcional insere o conceito como elemento integrador no contexto analisado, ou seja, permite que se identifique, na definição, a função/finalidade do conceito dentro da área em questão. Com isto, a definição do conceito permitiria através de suas características identificar novos conceitos relacionados, ou seja, identificar o conhecimento sobre o conceito em análise. Por exemplo, seja a definição de mitocôndria: "***é uma*** organela semi-autônoma, auto replicante, que ***ocorre em*** vários números, formas e tamanhos no citoplasma de virtualmente todas as células eucarióticas. Ela é notavelmente ***o local da*** respiração dos tecidos". Uma paráfrase da definição permite-nos destacar relações (*é um*, *ocorre em*, *local da*) que podem ser explicitadas: (i) mitocôndria ***é uma*** organela semi-autônoma e replicante; (ii) mitocôndria ***ocorre no*** citoplasma; (iii) mitocôndria ***local da*** respiração celular.

4.4. REUSO INTEGRADO

No reuso integrado o usuário final interage com as ontologias reutilizadas e o registro do conceito. O objetivo é municiá-lo de informações para a tomada de decisão em relação à sua finalidade de uso das ontologias que, no experimento em questão, é a anotação de genomas.

O reuso se dá com o apoio de ferramentas de software que devem fazer uso das informações que o registro de conceito agrega às ontologias e tem seu foco no uso de um conceito por vez. Por exemplo, o usuário final busca por um termo adequado dentre as ontologias reutilizadas, no ambiente proposto. O ambiente mostra o termo nas várias ontologias onde ele aparece, os possíveis relacionamentos encontrados desse termo dentre as ontologias reutilizadas, e as demais informações contidas no registro de conceito, para cada termo selecionado.

Nesse cenário, o auxílio à tomada de decisão é dado de diversas maneiras, por exemplo: (i) é possível identificar termos semelhantes em duas ou mais ontologias ao termo eleito como conceito, sendo que um desses termos possui uma descrição textual mais precisa do que a o termo eleito. Nesse caso o usuário pode registrar uma observação nesse sentido em relação ao conceito e, posteriormente, encaminhá-la aos fóruns de discussão da comunidade que usa essas ontologias; (ii) o usuário percebe, através dos conceitos relacionados em outras ontologias, que seu escopo de uso das ontologias agora está ampliado para contemplar facetas que antes ele teria de buscar separadamente nas ontologias relacionadas, ou até mesmo poderia considerar não disponíveis; (iii) o usuário percebe que um determinado conceito não existe nas ontologias selecionadas e pode incluí-lo no registro de conceito para uma pesquisa posterior em outras ontologias ou para submetê-lo como sugestão à comunidade global que usa as ontologias. Enquanto isso, ele pode se beneficiar do uso de um termo padronizado, armazenado no registro de conceito, que representa o conceito sugerido.

5. CONCLUSÕES

O grande volume de dados em diferentes áreas temáticas, cuja disseminação é facilitada pela implementação das Tecnologias de Informação e Comunicação e pelo avanço do conhecimento, tem gerado oportunidades para a atuação do profissional da informação, em uma vivência interdisciplinar em diferentes domínios. Em especial, em domínios científicos como a Biomedicina, a complexidade e a necessidade de compartilhar informação tornam crítica a adoção de um padrão terminológico. Além disso, à medida que os dados vão se acumulando, a possibilidade de seu tratamento com vistas à descoberta de conhecimento faz com que as ontologias venham ocupando um papel de destaque nesse cenário.

No domínio da Biomedicina, foco de nosso estudo, observamos a necessidade da integração desses vocabulários através de uma proposta que contemple não só os aspectos do tratamento computacional da sua compatibilização, mas também os aspectos do contexto onde se inserem as necessidades e objetivos do usuário final, trazendo-o para o centro do tratamento informacional. Nesse cenário, nossa investigação tem apontado para a importância do aporte teórico dos estudos em Compatibilização de Linguagens, na Ciência da Informação e na Ciência da Computação.

No estágio atual desta pesquisa, os testes estão sendo conduzidos de forma semi-automática em um conjunto restrito de 29 termos a partir de uma amostra selecionada da GO, que possui semelhança no que se refere à forma do termo (designação) com outras ontologias da OBO. Nos 29 termos analisados pudemos encontrar vários dos problemas de compatibilidade citados no tratamento de conflitos.

Esses problemas têm sido usados como subsídio para melhoria da precisão semântica das ontologias em nossos estudos.

Para apoiar essa tarefa foram desenvolvidos dois aplicativos de software protótipos, para mapear e manipular as ontologias mapeadas, com o foco no seu reuso integrado.

Trabalhos complementares estão sendo conduzidos para explorar a descoberta de novas relações entre as ontologias explorando a definição textual de seus termos.

REFERÊNCIAS

BYRNE, C. C.; McCRACKEN, S. A. "An adaptive thesaurus employing semantic distance, relational inheritance and nominal compound interpretation for linguistic support or information retrieval". *Journal of Information Science,* v. 25, n. 2, p. 113-131, 1999

CAMPOS, M. L. A. " Modelização de Domínios de Conhecimento: uma investigação de princípios fundamentais". *Ciência da Informação,* Brasília, v. 33, n. 1, p. 22-32, 2004.

CROSS, V., WANG, Y. "Semantic Relatedness Measures in Ontologies Using Information Content and Fuzzy Set Theory." In: *proceedings of the 14th IEEE International Conference on Fuzzy Systems*, p. 114-119, *2005*

DAHLBERG, I. "A Referent-oriented analytical concept theory of interconcept".

International Classification, Frankfurt, v.5, n.3, p.142-150, 1978b.

DAHLBERG, I. "Conceptual compatibility of ordering systems". *Internacional Classification,* v. 10, n. 2, p.5-8, 1983.

DAHLBERG, Ingetraut. *Ontical structures and universal classification*. Bangalore : Sarada Ranganthan Endowment, 1978. 64 p.

DAHLBERG, I. "Towards establishment of compatibility between indexing languages." *Internacional Classification,* v. 8, n. 2, p. 88-91, 1981.

DE BRUIJN, J., EHRIG, J., FEIER, C., MARTIN-RECUERDA, F., SCHARFFE, F., WEITEN, M.: "Ontology mediation, merging and aligning" , In: *Semantic Web Technologies.* Wiley, 2006.

DING, Y; FOO, S. "Ontology research and development. Part 1 - a review of ontology mapping and evolving". *Journal of Information Science,* v. 28, n. 2, p. 123–136, 2002.

EHRIG, M., HAASE, P., STOJANOVIC, N., HEFKE. M., "Similarity for Ontologies - a Comprehensive Framework". In *Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability,* Springer, December 2004.

EUZENAT, J., SHVAIKO, P.: *Ontology Matching.* Springer, 334p, 2007, ISBN: 9783-540-49611-3

FELICÍSSIMO, C.H.; BREITMAN, K. "Taxonomic ontology alignment – an implementation.". In: *Workshop De Engenharia De Requisitos,* Argentina, Tandil, p. 152-163, 2004.

GANGEMI, A.; STEVE, G.; GIANCOMELLI, F. *"ONIONS: An Ontological Methodology for Taxonomic Knowledge Integration", Workshop on Ontological Engineering,* Budapest, 1996.

GENE ONTOLOGY CONSORTIUM. "Creating the gene ontology resource: design and implementation". *Genome Research,* v. 11, n. 8, p. 1425-1433, 2001.

GLUSHKOV, V.M.; SKOROKHOD'KO, E.F.; STRONGNII, A. A." Evaluation of the degree of compatibility of information retrieval languages of document retrieval systems". *Automatic. Documentation & Mathematical Linguistics,* v. 12, n.1, p. 18-26, 1978

GRENON, P., SMITH, B., GOLDBERG, L., "Biodynamic Ontology: Applying BFO in the Biomedical Domain". *Ontologies in Medicine, Amsterdam*: IOS Press, p. 20–38, 2004.

GRUBER, T.R. "A translation approach to portable ontology specifications." *Knowledge Acquisition,* v. 5, p. 199-220, 1993.

GUARINO, N. "Formal ontology and information systems". In*: Proceedings of the Formal Ontologies in Information Systems, Trento, Italy* : IOS Press,. p. 3-15. 1998.

HEIDORN, P. B.; PALMER, C. L.; WRIGHT, D. "Biological information specialists for biological informatics." *Journal of biomedical discovery and collaboration,* v. 2, n. 1, 2007.

HJØRLAND, B. "Domain analysis in information science: eleven approaches – traditional as well as innovative." *Journal of Documentation,* v. 58, n. 4, p. 422– 462, 2002.

KALFOGLOU, Y, SCHORLEMMER, M. "Ontology Mapping: The State of the Art." *The Knowledge Engineering Review* 18(1): p.1-31, 2003.

LATOUR, B., *Vida de laboratório. Rio de Janeiro, Relume-Dumará, 1997*

MAEDCHE, A., STAAB, S.: "Measuring Similarity between Ontologies." In. *Proceedings of the 13th International Conference of Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web,* p. 251-263, 2002.

MENDES, P. N. *"Uma Abordagem para Construção e Uso no Suporte à Integração e Análise de Dados Genômicos"*. Dissertação de Mestrado – NCE - UFRJ, Rio de Janeiro, 2005.

MILLER, G., "Wordnet: An On-line Lexical Database". *International Journal of Lexicography,* 3(4) : p. 235—312, 1990.

NEVILLE, H. H. "Thesaurus reconciliation". *Aslib Proc.,* v.11, n.24, p. 620-6, nov. 1972.

NOY, N. F.; MUSEN, M. A. *"PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment." 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*", *2005*. [consulta: 17 maio 2008]. Disponível em: <http://obo.sourceforge.net.>.

PINTO, H.S., MARTINS, J. P., "A Methodology for Ontology Integration", *Proceedings of First International Conference on Knowledge Capture,* Victoria, B.C., Canada, ACM Press, 2001.

RADA, R., MARTIN, B. "Augmenting Thesauri for Information Systems", *ACM Transactions on Office Information Systems*, 5, 4, pp. 378-392, 1987.

RANGANATHAN, S. R. *Prolegomena to Library Classification.* New York: Asia Publishing House, 1967.

SABOU, M., D'AQUIN, M., MOTTA, E. "Using the Semantic Web as Background Knowledge in Ontology Mapping", *Ontology Mapping Workshop,* ISWC06, 2006.

SHATFORD, S. *Analyzing the Subject of a Picture: A Theoretical Approach*. Physical Sciences and Technology Libraries, University of California, Los Angeles, 1986.

SMITH, B. *"Ontology and information systems", 2002. [Consulta: 26 maio 2008]. Disponível em: < http://ontology.buffalo.edu/ontology(PIC).pdf>*

SOERGEL, D. "Compatibility of vocabularies." In*: Proceedings of conference on conceptual and terminological analysis in the social sciences.* Frankfurt, Verlag, p. 209-223, 1982.

SU, X. *Semantic Enrichment for Ontology Mapping*. PhD thesis Dept. of Computer and Information Science, Norwegian University of Science and Technology, 2004. ISBN 82-471-6453-1.