RECIIS
Revista Eletrônica de Comunicação Informação & Inovação em Saúde
[www.reciis.cict.fiocruz.br]
ISSN 1981-6278

Artigos originais

Aspectos metodológicos no reuso de ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos

DOI: 10.3395/reciis.v3i1.243pt

Maria Luiza de Almeida Campos
Instituto de Artes e Comunicação Social, Universidade Federal Fluminense, Niterói, Brasil
marialuizalmeida@gmail.com

Maria Luiza Machado Campos
Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
mluiza@pq.cnpq.br

Alberto M. R. Dávila
Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil
davila@fiocruz.br

Hagar Espanha Gomes
Universidade Federal Fluminense, Niterói, Brasil
hagar.espanha@terra.com.br

Linair Maria Campos
Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal Fluminense, Niterói, Brasil
linair@hotmail.com

Laura Lira
Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal Fluminense, Niterói, Brasil
llira@gbl.com.br

Resumo
Nos últimos anos tem havido um impulso no número de ontologias produzidas, refletindo um contínuo amadurecimento dos esforços de desenvolvimento de vocabulários, em especial na área de Biomedicina, caracterizada como um domínio inter e multidisciplinar e de temática complexa. Entretanto, apesar de haver propostas metodológicas e de melhores práticas sobre como organizar a estrutura terminológica de ontologias e de suas relações, pouco se explica sobre os métodos adotados para o levantamento terminológico do seu domínio e da delimitação de seu escopo, especialmente considerando o reuso de ontologias. O objetivo desse trabalho é apresentar as bases da Ciência da Informação e da Computação para atividades de reuso em ontologias, como um passo metodológico para a aquisição de conhecimento, visando possibilitar mecanismos para o mapeamento e alinhamento de termos em ontologias no domínio dos Tripanosomatídeos.

Palavras-chave
reuso de ontologia; alinhamento de ontologia; compatibilização de linguagem; tripanosomatídeos; aquisição de conhecimento

Introdução

No campo da genômica, iniciativas da comunidade científica internacional, nos últimos anos, levaram a um crescimento explosivo de informações biológicas, as quais vêm sendo geradas todos os dias de forma contínua (HGP 2003). A preocupação inicial, então, foi a criação e manutenção de bancos de dados para armazenar informação biológica. Conforme as bases de dados genômicas vão sendo preenchidas e os genomas seqüenciados, o foco das pesquisas começa 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. Torna-se fundamental a interligação entre os dados obtidos pelos diversos projetos de pesquisa ao redor do mundo sobre o inter-relacionamento de enzimas, genes, componentes químicos, doenças, espécies, tipos de células, órgãos etc. (Mendes 2005). Para que estas 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, a integração e reuso dos mesmos.

Desta forma, é importante considerar a relevância da gerência, descrição e organização dos recursos científicos em meio digital para a pesquisa em Bioinformática. Cabe observar que a Bioinformática é uma área interdisciplinar na qual Biologia, Ciência da Computação e Tecnologia da Informação fazem parte e cujo objetivo é permitir a descoberta de novas introspecções biológicas, assim como criar uma perspectiva global de que os princípios unificados da biologia podem ser discernidos (Belloze 2007).

A grande quantidade de dados que está sendo acumulada nos diferentes bancos de dados ao redor do mundo precisa, a partir das informações genômicas disponíveis, ser anotada e interpretada. Para este fim, é necessário que os diversos projetos interessados em trocar e integrar informações descrevam seus dados de forma padronizada, de modo a possibilitar com consistência a recuperação de informações. Ontologias assumem papel fundamental nesta integração, viabilizando a interoperabilidade semântica de sistemas distribuídos heterogêneos, como é o caso de esforços que reúnem consórcios internacionais (Campos 2007).

A Biblioteca de Ontologias OBO (Open Biological Ontologies) (Obo 2005) é um repositório de terminologias desenvolvido para uso compartilhado entre diversos domínios biológicos e médicos. Apesar de se denominar um repositório de *ontologias*, na verdade, os vocabulários existentes podem ser definidos de diversas formas, como: vocabulários controlados, glossários e propriamente ontologias. Além disto, alguns vocabulários objetivam ser genéricos a ponto de serem aplicáveis a quaisquer organismos, enquanto outros contem termos específicos a grupos taxonômicos tais como moscas, fungos, leveduras ou peixes. Dentre os mais difundidos vocabulários componentes da OBO, podemos destacar a Gene Ontology (GO) (Gene Ontology Consortium 2001). A GO compreende termos referentes a três grandes categorias: componentes celulares, processos biológicos e funções moleculares, de maneira independente de espécies de organismos (Ashburner 2002).

No Brasil, especificamente nas atividades da área de aplicações científicas genômicas, vem sendo desenvolvido o projeto "Genoma e Transcriptoma comparativo: um consórcio de Bioinformática para o desenvolvimento de uma plataforma Web e bancos de dados integrados", coordenado pela Fiocruz. Este projeto tem como um dos principais objetivos prover um ambiente que possa oferecer informação semântica sobre recursos científicos, como dados e programas na área de Bioinformática, e possibilitar o uso destes recursos de forma conjunta pela comunidade científica interessada. A GO tem sido utilizada para as anotações em seu banco de dados.

Para a implementação deste ambiente, foi formado um consórcio envolvendo a Fiocruz e as Universidades Federais do Rio de Janeiro e Santa Catarina visando o desenvolvimento de um portal de Bioinformática e uma plataforma web integrada para análises de genomas e transcriptomas. O desenvolvimento da capacidade e infra-estrutura em Bioinformática no Brasil é estratégico e conseqüentemente de grande relevância para colaboração com as diferentes iniciativas dos projetos genoma tanto no Brasil como no exterior. Desta forma, para auxiliar, otimizar e disseminar as pesquisas, está sendo implementada progressivamente uma plataforma denominada de BiowebDB, fruto de um consórcio de mesmo nome, e que se encontra disponível publicamente em: http://www.biowebdb.org.

O Consórcio BioWebDB, 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 e banco de dados genômicos. A Genômica Comparativa compreende a análise e comparação de genomas de diferentes espécies, com o objetivo de atingir um melhor entendimento de como as espécies evoluíram ou de determinar a função de genes e regiões não codificantes do genoma através dessas comparações. Muito do que existe de informação sobre genes humanos pode ser descoberto graças à análise de seus correlatos em organismos-modelo mais simples, tais como camundongo (HGP 2003).

As pesquisas do grupo encontram-se concentradas em três principais focos: no desenvolvimento de ferramentas de Bioinformática para análise de genomas, análise dos genomas de tripanosomatídeos, e desenvolvimento de ontologias e compatibilização de linguagens. A iniciativa do consórcio é construir plataformas flexíveis, integrados e amigáveis, capazes de serem compartilhados com diferentes conjuntos de dados e projetos de genoma. Neste sentido, as ontologias ganham uma importância fundamental para garantir a harmonização semântica e a recuperação de informações.

O estudo que ali estamos desenvolvendo, já aponta para alguns resultados que possibilitam afirmar que até o momento não se identificam, a nível nacional e internacional, ontologias desenvolvidas dentro do recorte conceitual específico, ou seja, de tripanosomatídeos para atender as demandas dos grupos coordenados pela Fiocruz. Apesar

dos esforços internacionais, a Gene Ontology não possui classes de conceitos que venham atender plenamente as pesquisas desenvolvidas no Brasil. Em alguns casos é necessário investigar a harmonização existente entre termos e o seu conteúdo conceitual. Nesta medida, ainda como uma proposta do consórcio OBO, vem sendo incentivada a elaboração de recortes específicos da GO, chamados de GO Slims[1], cuja finalidade é fornecer sub-conjuntos da GO, muitas vezes com hierarquias menos profundas, voltados para organismos específicos.

No entanto, apesar da difusão de linguagens e ferramentas para a representação e construção de ontologias, as metodologias que as embasam resultam de pouca utilidade, pois em geral ainda não contemplam diretrizes satisfatórias nem para identificação dos conceitos e seus relacionamentos, nem tampouco para a criação de definições sistemáticas associadas a esses conceitos. Por conseqüência, as ferramentas têm pouco a contribuir no sentido de orientação do usuário no processo de construção da ontologia, assim como em diretivas para a construção de ontologias de qualidade (Gangemi et al. 1996, Fernández et al. 1997).

Neste artigo pretendemos problematizar as questões que envolvem o reuso de ontologias, como um passo metodológico para a aquisição de conhecimento em ontologias, visando possibilitar mecanismos para o mapeamento e alinhamento de termos em ontologias no domínio dos tripanosomatídeos.

Neste sentido o trabalho encontra-se assim organizado: na seção 2, tratamos dos aspectos básicos do reuso de ontologias; na seção 3, os trabalhos relacionados; na seção 4 detalhamos de forma preliminar nossa proposta para os aspectos metodológicos aplicados no reuso de ontologias; na seção 5 apresentamos uma discussão sobre nosso trabalho e as dificuldades encontradas; por fim, na seção 6, as considerações finais.

Reuso de ontologias

Como apresentado em diversos estudos, ontologia (Gruber 1993, Guarino 1993, 1998, Vickery 1997, Swartout & Tate 1999, Corazzon 2000, Smith 2002) como instrumento de representação de conhecimento, surge no âmbito da Inteligência artificial na década de 90. Para os sistemas de Inteligência Artificial, o que existe é o que pode ser representado. 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) "uma ontologia é uma especificação de uma conceituação".

Uma conceituação é uma abstração, uma visão simplificada do mundo que se representa para satisfazer um ou mais dos seguintes propósitos: "permitir que múltiplos agentes compartilhem seus conhecimentos; ajudar as pessoas a compreender melhor certa área de conhecimento; ajudar pessoas a atingir um consenso no seu entendimento sobre uma área de conhecimento" (Smith apud Falbo 1998). Em Lógica, uma conceituação identifica o objeto e relações que existem no universo lógico (Weinstein 1998).

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. O segundo caso é a abordagem que utilizamos, a qual é denominada de *alinhamento* de ontologias.

O alinhamento difere da junção e integração 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, porém gera 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, Aleksovski et al. 2006, 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. Um dos aspectos do mapeamento é a questão de como achar os candidatos. Para detalhes sobre essas questões, De Bruijin et al. 2006, apresentam um consistente levantamento sobre os tipos de conflitos ao se mapear ontologias.

Um outro aspecto para a obtenção de correspondências diz respeito ao tipo de técnica utilizada para estimar os candidatos. Esta pode se basear, dentre outros: (i) na semelhança dos nomes dos termos; (ii) na estrutura da ontologia, como, por exemplo, levando 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) na adição de conhecimento adicional, como, por exemplo, nas informações de uma terceira ontologia ou vocabulário que possua uma hierarquia de

conceitos, como a Wordnet (Miller 1990), que pode ser utilizada, por exemplo, para procura de sinônimos ou para confronto da distância do posicionamento dos termos das ontologias sendo mapeadas em relação a essa terceira ontologia (Reynaud & Safar 2007, Sabou et al. 2006).

Trabalhos relacionados ao reuso de ontologias

A literatura sobre reuso de ontologias 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).

Alguns autores chegam a propor tarefas mais gerais que são necessárias no processo de reuso. Gangemi et al. 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, em uma visão mais abrangente, sugerem que o reuso começa na seleção de ontologias a serem reutilizadas. Entretanto, não fornecem muitos detalhes sobre como devem se dar essas tarefas.

Os nossos estudos têm apontado para a importância da investigação no âmbito dos estudos em Compatibilização de Linguagens no domínio de Ciência da Informação. Consideramos que a partir deles possamos obter diretrizes teóricas e metodológicas para o reuso em ontologia (Campos 2005).

Aspectos semânticos do reuso ligados à compatibilização de vocabulários

Um dos aspectos do reuso é a compatibilidade entre os vocabulários reutilizados. Cabe aqui ressaltar que o termo compatibilidade no âmbito da Ciência da Computação tem definição bastante específica. Refere-se à capacidade dos computadores de vários tipos de utilizar programas escritos para outros sem conversão para outras linguagens de máquina. 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, com teóricos como Soergel (1982), Dalhberg (1981), Neville (1970,1972) e Glushkov (1978), (Campos 2006).

Para Glushkov et al. (1978) compatibilidade é a medida de similaridade entre duas linguagens, onde se introduz o conceito de graus de compatibilidade e estabelecem a distinção entre compatibilidade em plano semântico e no plano linguístico.

Dos métodos de compatibilização e conversão de linguagens, baseados na integração de vocabulários, dois se destacam sobremaneira. São o método de reconciliação de tesauros proposto por Neville (1970, 1972) e a matriz de compatibilização conceitual proposta por Dahlberg (1981, 1983).

O método de Neville baseia-se no princípio que se deve 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, baseado na codificação numérica de conceitos através do qual se torna possível o estabelecimento da equivalência conceitual de descritores de diferentes linguagens.

O método proposto por Dahlberg (1983) baseia-se na 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 vistas semântico e estrutural. A compatibilidade entre linguagens, segundo Dalhberg, compreende três fases, são elas: 1. a coincidência conceitual – quando dois conceitos combinam suas características – grau de equivalência; 2. Correspondência conceitual - dois conceitos combinam a maior parte de suas características – similaridade; 3. correlação conceitual - dois conceitos são correlacionados através de símbolos matemáticos, estabelecendo uma medida de correlação.

A compatibilização, entretanto, pressupõe que os vocabulários devem possuir algum grau de compatibilidade, e quanto mais compatíveis, mais fácil e precisa é a sua compatibilização. Para serem mais compatíveis os vocabulários devem, idealmente, seguir normas que forneçam diretrizes para a sua construção mais uniforme e padronizada. Lancaster (1986) já observava essa questão no âmbito da construção de tesauros:

> "As normas, ao promoverem a compatibilidade estrutural dos vocabulários, facilitam a conversão de um vocabulário para outro. Assim, dois tesauros seguindo as normas ISO para construção de tesauros, são provavelmente mais facilmente reconciliados do que dois construídos com princípios diferentes. Ainda mais, tais normas promovem a compatibilidade de um modo geral: Uma vez familiarizado com um tesauro, seria mais fácil para um usuário de um serviço de informação converter para outro tesauro construído de acordo com as mesmas convenções." (Lancaster 1986, p 212).

É importante observar que na grande maioria das vezes as propostas de alinhamento exploram compatibilização de termos com significado semelhante, assumindo que é preciso conviver com diferentes vocabulários que tratam de temáticas com algum grau de sobreposição. Essa forma de conceber as ontologias como vocabulários expressando diferentes visões de um mesmo domínio, entretanto, não é consensual. Em especial na área Biomédica, onde os vocabulários são de temática complexa.

Alguns autores, tais como N. Guarino e Barry Smith apresentam propostas um pouco diferentes, embora ambas sejam voltadas para a padronização de ontologias a partir do estudo da categorização dos seus conceitos e relações.

Guarino (1998a) possui, dentre outros, estudos que exploram a natureza semântica e formal dos conceitos de uma ontologia. Na prática, a *Ontologia Formal*, de Guarino, pode ser entendida como a teoria das distinções a priori sobre: as entidades do mundo (objetos físicos, eventos, regiões, quantidades de matéria); as categorias de meta-nível para modelar o mundo (conceitos, propriedades, qualidades, estados, papéis e partes). Guarino, entretanto, admite que sejam criadas várias visões não necessariamente complementares de um mesmo domínio, que denomina de "mundos possíveis".

Barry Smith (Smith et al. 2007), por sua vez, busca inspiração na Teoria de Classes de Aristóteles para propor um conjunto de axiomas e definições para aplicação no domínio da Biomedicina, desenvolvido de forma colaborativa. Embora a visão de Smith de categorização da ontologia seja filosoficamente próxima à de Guarino (Bateman & Farrar 2004), Smith, em oposição a Guarino, defende a idéia de que existe apenas um "mundo possível", embora com diferentes visões, ortogonais, que se complementam. Para Smith as ontologias:

> "(i) devem ser desenvolvidas em um esforço colaborativo, (ii) usam relações comuns que são definidas de forma não ambígüa, (iii) ... (iv) têm uma temática claramente delimitada (de modo que uma ontologia voltada para componentes celulares, por exemplo, não inclua termos como 'banco de dados' ou 'inteiro').."(Smith et al. 2007, p.2).

Além de investigar as abordagens de teóricos como Smith e Guarino, nossa pesquisa tem se apoiado em estudos desenvolvidos no campo da compatibilização de linguagens, no âmbito da Ciência da Informação. Especialmente nas teorias ligadas mais especificamente à representação de sistemas de conceitos, onde existe uma base teórica sólida para a elaboração de linguagens de vertente européia, que irá possibilitar uma base semântica para a integração, como: a Teoria da Classificação Facetada de S. R. Ranganathan (Ranganathan 1967) e a Teoria do Conceito de I. Dahlberg (Dahlberg 1978 a, b, 1983), que possibilitam a representação de domínios de conhecimento. Pelo enfoque abordado neste artigo, não detalharemos essas teorias, porém trataremos brevemente da contribuição de Ranganathan, uma vez que delas fazemos uso na fase atual de nosso trabalho, conforme ilustrado na seção 4.

Ranganthan elabora uma série de princípios que visam a permitir que os conceitos de um domínio de saber possam ser estruturados de forma sistêmica, isto é, os conceitos se organizam em renques e cadeias, estas estruturadas em classes abrangentes, que são as facetas, e estas últimas dentro de uma dada categoria fundamental. A reunião de todas as categorias forma um sistema de conceitos de uma dada área de assunto e cada conceito no interior da categoria é também a manifestação dessa categoria (Campos 2001). A Categorização é um processo que requer pensar o domínio de forma dedutiva, ou seja, determinar as classes de maior abrangência dentro da temática escolhida. O exercício de categorização pode tornar claro o domínio temático da ontologia e, como conseqüência, estabelece as bases para seleção dos termos, nas fontes de onde eles serão retirados.

Neste espaço é que a base onde se fundamenta sua teoria pode auxiliar no recorte de domínio para a elaboração de ontologias e fundamentalmente para a construção de modelos conceituais. O seu postulado das [Meta]Categorias, de especial interesse para nosso estudo, propõe a existência de cinco categorias fundamentais, que podem ser usadas para se recortar universos de assunto em classes abrangentes. Independentes de quais categorias são usadas para se pensar à estruturação de um domínio (cinco, menos ou mais), a idéia de que estas agrupam conceitos, como propõe Ranganathan, é um fator importante a se considerar quando da compatibilização de vocabulários, uma vez que elas permitem aumentar a semântica da natureza das classes. Esta perspectiva está sendo explorada em nossa experimentação em sua fase inicial. No futuro esperamos explorar as outras contribuições da CI referidas anteriormente nesta seção.

Conforme podemos observar, a organização de domínios de conhecimento tem recebido destaque tanto na Ciência da Informação quanto na Ciência da Computação, de forma bastante independente e, por vezes, pontual em certos aspectos, como, por exemplo, a organização de conceitos hierarquicamente, ou a eficiência de algoritmos computacionais. Nossa proposta pretende aproximar essas áreas e ampliar e integrar, quando possível e pertinente, a discussão das propostas de organização de domínios no escopo do reuso de ontologias.

Neste cenário, a partir da revisão da literatura, parece haver uma carência de propostas que tratem de maneira abrangente e detalhada de questões que antecedem e fundamentam o reuso em si das ontologias, situando-as em um contexto que permita compreender a sua origem, motivação, objetivo e cenários de aplicação. Critérios para a escolha das ontologias vão além da identificação de suas características a serem analisadas. Estes devem considerar não só os princípios que devem nortear tal análise, como também os princípios para delinear o contexto onde as ontologias vão ser reaproveitadas, tanto do ponto de vista da sua aplicação imediata, quanto do ambiente onde se inserem. Nossa hipótese é que a partir da identificação e do detalhamento de tais princípios, o reuso se dá de forma mais consistente e precisa.

Compatibilização de ontologias: aplicação no domínio dos tripanosomatídeos

A concretização de nossa proposta se dá, como mencionado anteriormente, no âmbito dos projetos do Consórcio BioWebDB, conjugando esforços de natureza teórica e experimental, reunindo uma equipe de natureza interdisciplinar, contando com pesquisadores de diferentes instituições[2]. Dentro desta perspectiva, nesta seção trataremos de abordar os primeiros experimentos em torno da compatibilidade de ontologias a partir do conceito de reuso. Até o momento, em nosso experimento tratamos de dois enfoques principais, quais sejam, a metodologia utilizada para compor a amostra de termos e a abordagem de reuso adotado para aplicação na amostra selecionada.

Levantamento do vocabulário no domínio dos tripanosomatídeos

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, Lancaster 1986, Dahlberg 1978b, Hjorland 2002). Estes estudos fornecem diretrizes sistemáticas que têm sido investigadas, no contexto deste trabalho, para uma análise preliminar do domínio. Através do apoio destes aportes teóricos e de outros das Ciências Sociais (Latour 1997), elaboramos um primeiro esboço dos agrupamentos temáticos do domínio dos Tripanosomatideos no Laboratório de Biologia Molecular de Tripanosomatídeos e Flebotomíneos: do Instituto Oswaldo Cruz (IOC).

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 estes 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 vimos participando de uma série de seminários e entrevistas, que ajudam a compreender melhor esse domínio.

Hjorland (2002), ainda, 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, estudo do usuário, dentre outros.

Tendo em mente estas perspectivas, a análise do domínio do pesquisador seguiu, em um primeiro momento, um critério de mapeamento tanto das atividades desenvolvidas no Laboratório quanto da literatura, visando identificar, por um lado, um conjunto de ontologias onde ferramentas de reuso poderiam ser aplicadas e, por outro, um conjunto de termos como base de amostra para atividades de compatibilização, como veremos adiante.

A partir da literatura resultante das pesquisas realizadas no âmbito do Laboratorio, foi feito um levantamento das temáticas e das ontologias de interesse (além da GO). A princípio, dez grandes agrupamentos temáticos foram identificados: Protistas, Biologia funcional e de Sistemas, Biologia molecular e Genômica, Genética molecular evolutiva, Genômica comparativa, Filogenia, Bioinformática, Doenças, Metagenômica, Alvos para fármacos, cada um destes com sub-grupamentos que estamos detalhando e validando no presente momento[3].

Assim, foram mapeadas as ontologias no escopo da OBO relacionadas com os agrupamentos temáticos. Quanto às ontologias de interesse chegou-se a um conjunto de sete ontologias: NCBI organismal classification, Pathway, Sequence types and features (SO), Brenda tissue/enzyme source, Event (INOH pathwayontology), Multiple alignment e System biology (OBO 2005). Estas serão utilizadas como domínio para que possamos identificar classes no âmbito do domínio dos Tripanosomatídeos.

Por outro lado, um conjunto de 800 termos, resultantes das anotações genômicas existentes no sistema GARSA (Davila et al. 2005)[4], anotados com base na Gene Ontology (GO) e produto de pesquisa (Wagner 2006), no âmbito da genômica funcional de tripanosomatídeos, em particular a da espécie T. rangeli, foram utilizados para comparação com as ontologias da OBO selecionadas de modo a obtermos várias hierarquias de termos pais e filhos para cada termo encontrado, com as suas respectivas definições e seus relacionamentos partitivos, quando existentes. Para isso foi desenvolvido um aplicativo de software. Esse aplicativo é desenvolvido não só para extrair, mas também para converter a linguagem das ontologias utilizadas (originalmente em formato OBO) para a linguagem OWL (Web Ontology Language) (OWL 2008) de modo a facilitar futuras inferências e manipulações computacionais, uma vez que este material com as hierarquias das ontologias será utilizado como amostra para os experimentos com reuso.

Abordagem adotada para o reuso

A escolha da abordagem para reuso depende, dentre outros fatores, do objetivo que se pretende atingir e do contexto onde se insere o seu uso. No caso de nosso cenário experimental, o objetivo é a descrição de seqüências genômicas de tripanosomatídeos, dentro de uma visão integrada do genoma, transcriptoma, proteoma e metaboloma desses organismos. Para isso, há que se considerar os seguintes fatores de seu contexto de uso: (i) o vocabulário largamente utilizado em Biomedicina, a GO, deve ser não apenas reduzida em seu escopo para os Tripanosomatídeos, como também complementada com outros que dizem respeito a aspectos não cobertos por ela, como, por exemplo, vias metabólicas e doenças; (ii) a descrição dessas seqüências deve apontar, de algum modo, para vocabulários padronizados da área, em especial a GO, devido ao emprego hegemônico desses vocabulários na anotação de recursos genômicos; (iii) o grupo de pesquisa não pode arcar com o ônus de atualização das ontologias criadas, uma vez que seus recursos são escassos.

Cabe ressaltar que, apesar de existirem esforços para a reformulação das ontologias da OBO, visando a sua fatoração em ontologias ortogonais, bem definidas e organizadas, esta ainda não é a realidade atual. Desta forma, enquanto essa iniciativa não se consolida, é importante lidar com a sobreposição de temáticas e conceitos semelhantes existindo em ontologias de temáticas diversas e com definições distintas.

Ao levar em conta os fatores acima, concluímos que em nosso cenário de estudo o processo de alinhamento é o mais indicado. A estratégia metodológica adotada para o alinhamento das ontologias selecionadas no item 4.1, baseia-se no critério da compatibilização semântica apoiada por conhecimento adicional, este, obtido em um primeiro momento, apenas a partir do estudo e identificação da natureza das classes de primeiro nível das ontologias. Este estudo se dá sob a perspectiva de

categorias fundamentais, e apóia-se na Teoria da Classificação (Ranganthan 1967).

Em relação à realização efetiva do alinhamento da ontologia, devido à complexidade da tarefa e da possibilidade de automação de certas atividades, destaca-se a importância do apoio de ferramentas de software, como, por exemplo, a Prompt (Noy & Musen 2000), Chimera (Mcguinness et al. 2000) ou Fca-Merge (Stumme & Madche 2001). Em especial em relação à tarefa de encontrar os termos candidatos (correspondência) para mapeamento.

Neste sentido, pretendemos verificar se a aplicação dos princípios metodológicos propostos contribuem para a melhoria da *precisão* de ferramentas de software na obtenção de termos de interesse para a nossa área de experimentação. Para isso, trabalhamos especificamente na adaptação de uma ferramenta de software desenvolvido como fruto de um projeto de graduação do curso de Ciência da Computação da UFRJ, cujo objetivo é o alinhamento de ontologias através de um algoritmo que explora a sua estrutura hierárquica e as propriedades de suas classes, em relação à semelhança de seus nomes (SILVA, 2008). Em nossa adaptação, além da estrutura da ontologia e da semelhança dos nomes, propomos considerar a natureza semântica das classes e propriedades.

Resultados preliminares

No estágio atual, nossos testes estão sendo conduzidos de forma semi-automática, em um conjunto restrito de 28 termos a partir da amostra selecionada aleatoriamente.

Para cada ontologia da OBO, um recorte é efetuado gerando ontologias com as hierarquias ascendentes e descendentes dos termos encontrados. Cada uma dessas ontologias é então mapeada com um subconjunto da GO que contém as hierarquias dos 28 termos selecionados. O mapeamento é efetuado com a ajuda da ferramenta Prompt. O apoio de ferramentas é fundamental em Biomedicina, devido ao grande número de termos de suas ontologias (algumas com mais de dezenove mil).

Cada sugestão de mapeamento feita pela ferramenta é então analisada manualmente para avaliar três aspectos: similaridade na designação dos termos, similaridade semântica indicando conceitos de natureza semelhante (relacionados logicamente), relacionamento indicando conceitos que não são semelhantes, mas que podem estar associados através de relações categoriais (lógicas) relevantes para o domínio. Neste último caso, buscamos, no momento, avaliar a complexidade e viabilidade de efetuar essa tarefa manualmente.

O objetivo de tal experimento é identificar um conjunto ideal de mapeamentos sugeridos com precisão, ou seja, com um máximo de sugestões que possam ser aproveitadas. A metodologia adotada visa aumentar a semântica das ontologias manipuladas. Cabe observar que no atual estágio de nossa pesquisa a técnica utilizada para estimar as correspondências dos termos candidatos ao mapeamento (vide Figura 1) apóia-se na semelhança da designação dos termos, na análise da estrutura da ontologia e do uso de conhecimento adicional, a ser incorporado através de uma ontologia formal de alto nível, elaborada especificamente para o domínio em questão.

Com base em uma análise preliminar, nossos resultados já sugerem um aumento de precisão no tratamento de falsos positivos, o que nos aproxima do conjunto ideal de mapeamentos desejados. Estes resultados ainda encontram-se em seus estágios iniciais, devendo ter seu escopo ampliado e revisado. Entretanto, podemos considerar que apontam para indícios promissores que confirmam a validade de nossa hipótese.

Como exemplo, podemos citar o mapeamento do conceito *excretion* (excreção) encontrados nas ontologias GO e Brenda. Na primeira, o termo diz respeito a um processo e significa "a eliminação por um organismo de dejetos que são resultados de uma atividade metabólica". Na segunda, diz respeito a um produto de uma atividade e significa "a matéria, tal como urina ou suor, que é excretada do sangue, tecidos ou órgãos". Quando mapeamos as duas ontologias através da ferramenta Prompt, esta sugere que os termos são semelhantes, mas de fato requerem uma análise semântica.

De maneira análoga, os termos *transporter*, da ontologia MoleculeRole (ramo da INOH) e *transport* da GO, também apresentam falsos positivos no mapeamento sugerido pela Prompt. *Transport*, na GO, é um processo definido como "processos pertinentes especificamente ao funcionamento de unidades vivas integradas: células, tecidos, órgãos e organismos.". *Transport* na MoleculeRole, por sua vez, é uma proteína definida como "que liga os solutos específicos a serem transportados e sofrem uma série de mudanças de conformação para transferir o soluto ligado (...)". A Figura 1 ilustra exemplos desse tipo de resultado, obtidos em nossas análises iniciais.

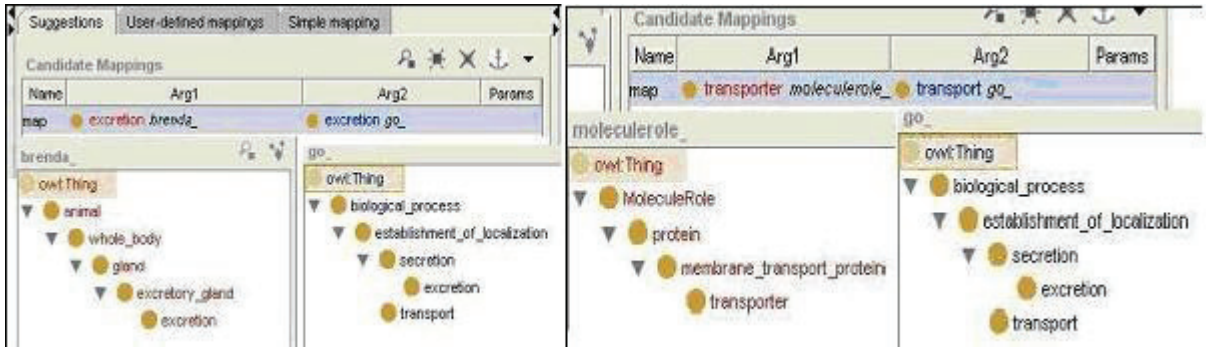

Figura 1 - Falsos positivos sugeridos para mapeamento pela ferramenta Prompt

Esses pares de termos, como podemos perceber, apesar de sua similaridade lingüística, denotam conceitos de naturezas distintas (diferentes categorias fundamentais), e, sendo assim, não deveriam ter sido sugeridos como candidatos para mapeamento por similaridade conceitual (relação de natureza lógica), como foi proposto pela ferramenta Prompt.

Por outro lado, a similaridade lingüística, quando confrontada com um conjunto de relações categoriais pré-definidas, que pode ser sugerida pela máquina para validação pelo homem, permite-nos identificar que os termos podem estar associados através de uma relação de natureza ôntica. No caso da Figura 2, fomos capazes de identificar o relacionamento entre os termos *excretion* (Brenda) e *excretion* (GO) através da relação categorial de *processo-produto*, ou seja, *excretion* (uma matéria na Brenda) *é produto de excretion* (uma atividade na GO). Da mesma forma, identificamos que *transporter* (uma proteína em MoleculeRole) *participa em transport* (um processo na GO).

Discussão

O domínio da Biomedicina, mesmo quando restrito aos estudos de espécies específicas dentro de um laboratório de pesquisas se revela complexo e desafiador.

Por um lado, há que se lidar com a dimensão humana, que se reflete na barreira das diferentes linguagens, dos conhecimentos que se complementam, tais como a do profissional de informática, do biólogo e do cientista da informação, cada um com um viés de pesquisa dentro do domínio, e possuindo diferentes graus de maturidade: desde os recém-graduados até os pesquisadores seniores com vasta experiência na área. Cada um destes tem uma visão diferente do domínio e essas visões devem ser conciliadas dentro de uma perspectiva mais ampla da Biomedicina, com aproveitamento de esforços já existentes.

Por outro lado, há que se lidar com a dimensão tecnológica, fundamental em uma área marcada pela complexidade e pela adoção de vocabulários da ordem de milhares de termos e com uma série de problemas, mas que são padrões de fato.

Neste cenário, nossos experimentos apontam para um enorme potencial de melhoria nas ontologias analisadas, as quais carecem de mecanismos que as integrem de maneira mais precisa, considerando não só os aspectos tecnológicos, da sua processabilidade pela máquina, mas também do seu entendimento pelo homem.

Nos 28 termos analisados pudemos encontrar vários problemas de compatibilidade, dentre eles: (i) definições de conceitos semelhantes em diferentes níveis de abstração; (ii) termos com denominação semelhante e significados distintos; (iii) termos que possuem relação entre si, porém sem que esta esteja explicitada, dentre outros. Estes problemas têm sido usados na nossa pesquisa como subsídios para melhoria da precisão semântica das ontologias, conforme exemplificado na Tabela 1.

Tabela 1 - Subsídios para melhoria da precisão semântica das ontologias analisadas

Natureza	Subsídios encontrados a partir da análise dos termos mapeados
(i)	O termo *transporter* na ontologia *system biology* é definido genericamente como: "entidade participante que facilita o movimento de outra entidade física de um subconjunto definido do ambiente físico (...) para outro" . Na ontologia MoleculeRole, a definição da entidade participante e entidade física é especificado para proteína e soluto, respectivamente. Ao confrontar essas duas definiçõe, podemos perceber que o uso de padrões definitórios pode trazer mais consistências para a formulação e compreensão dos conceitos. Por exemplo, a primeira definição citada acima poderia ser utilizada como um padrão definitório a ser seguido para outras mais específicas, como a segunda.
(ii)	As definições dos termos, confirmam, até o momento, os seguintes tipos de categorias fundamentais: processo biológico, função molecular, evento, componente biológico, componente químico, fenótipo.
(iii)	As definições dos termos apontam, até o momento, para os seguintes tipos de relação categorial (que não foram encontradas na relation ontology): processo-componente biológico, processo-evento disparador, processo-insumo, processo-produto, sendo que nestas duas últimas, o insumo e o produto são componentes químicos (orgânicos ou inorgânicos).

Considerações finais

A pesquisa em Biomedicina é marcada pelo grande volume de dados, pela complexidade temática e por um crescente número de vocabulários que buscam a descrição e organização dos recursos científicos relacionados.

Esses vocabulários têm sido construídos, na sua maioria, para atender a interesses que nem sempre atendem às necessidades da pesquisa no Brasil e, além disso, possuem problemas estruturais que sugerem a carência de metodologias voltadas para o seu desenvolvimento.

Entretanto, dada a alta complexidade do domínio, o alto custo envolvido na tarefa de construção de tais vocabulários, e a sua larga adoção pela comunidade biomédica, seu reuso tem de ser considerado na elaboração de vocabulários mais adequados à pesquisa nacional.

Nesse contexto, nossa proposta busca problematizar as questões que envolvem o reuso de ontologias, em particular as relacionadas ao mapeamento e alinhamento de termos em ontologias no domínio dos tripanosomatídeos.

Para isso, como ponto de partida, estamos conduzindo experimentos, voltados para a aquisição do conhecimento do domínio, que visam dar respaldo às bases teóricas que estamos investigando. Como resultado preliminar, destacamos um conjunto de 28 hierarquias de termos, com suas respectivas definições e relações partitivas, relevantes para a pesquisa do Laboratório de Biologia Molecular de Tripanosomatídeos e Flebotomíneos do IOC da Fiocruz. Essa amostragem, cujas temáticas se complementam e se sobrepõem em alguns aspectos, é importante instrumento de ensaios envolvendo questões de reuso de ontologias e está sendo explorada para fins de estudos de compatibilidade e da definição conceitual.

Uma exploração preliminar dessas hierarquias traz resultados que já apontam para a validade de nossa proposta de enriquecimento semântico das ontologias, a partir da identificação de categorias fundamentais, como importante fator para o aumento da precisão de ferramentas de software, cujo uso é fundamental em Biomedicina.

Trabalhos futuros, já esboçados, pretendem aprofundar a questão da análise do domínio através do tratamento semi-automático da literatura especializada da área, já levantada, da aplicação de outras contribuições da Ciência da Informação na área de compatibilização de vocabulários e do uso de ontologias de alto nível para se pensar as relações entre ontologias de temática complementar.

Notas

1. GO Slims são ontologias formadas a partir de um recorte da GO, contendo então um subconjunto de seus termos, e sendo geralmente utilizadas para a descrição de um determinado organismo ou de determinados aspectos biológicos apenas (por exemplo, apenas localizações celulares). Atualmente existem diversas GO Slims disponíveis, as quais podem ser obtidas a partir do site do consórcio da Gene Ontology.

2. Estes estudos são resultados preliminares de dois projetos de pesquisas, apoiados pelo CNPq, quais sejam: "Integração de Ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica " da área de Ciência da Informação; "Genoma e transcriptoma comparativo" da área da Ciência da Computação. Além dos projetos, eles são temáticas abordadas pelas pesquisa de dois doutorandos do Programa de Pós Graduação em Ciência da Informação UFF/IBICT. Em todas as pesquisas o campo empírico de atuação está vinculado aos estudos genômicos no âmbito do consórcio BioWeb DB.

3. Estamos, neste momento, testando algumas ferramentas de extração automática para levantar termos utilizando metodologia não manual.

4. Sistema desenvolvido na Fiocruz para análise e anotação de recursos genômicos.

Referências bibliográficas

Aleksovski Z, ten Kate W, van Harmelen F. Exploiting the Structure of Background Knowledge Used in Ontology Matching. In: Workshop on Ontology Matching at ISWC, 2006.

Ashburner M, Lewis S. On Ontologies for Biologists: the gene ontology – uncoupling the web. In: Silico Biology, Novartis Found Symposium, 2002, p. 66-83.

Bateman J, Farrar S. Towards a Generic Foundation for Spatial Ontology. In: Formal Ontology In Information Systems: Proceedings of the Third International Conference (FOIS-2004), 2004, p. 237-248.

Belloze KT. Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular. 2007. 147 f. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2007.

Campos ML. A Linguagem documentária: teorias que fundamentam sua elaboração. Niterói, RJ: Eduff, 2001.

Campos MLA. Integração de Ontologias: o domínio da bioinformática. RECIIS. 2007; 1:117-121.

Campos MLA. Integração de ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica. (Projeto de Pesquisa submetido ao CNPq no período de 2005 a 2008). Universidade Federal Fluminense- Departamento de Ciência da Informação, 2005a.

Campos MLA. Integração de ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica. In: VII Enancib, 2006, Marília. Anais... Marilia, 2006.

Campos MLM, Campos MLA, Campos LM. Web semântica e a gestão de conteúdos informacionais. In: Carlos H. Marcondes; Hélio Kuramoto; Lídia Brandão Toutain; Luís Sayão. (Org.). Bibliotecas digitais: saberes e práticas. Salvador, BA; Brasília: EDUFBA; IBICT, 2005, p. 55-75.

Corazzon R. Ontology: a resource guide for philosophers. 2000. Disponível em: <http://www.formalontology.it>. Acesso em: 1 jul. 2006.

Dahlberg I. A Referent-oriented analytical concept theory of interconcept. International Classification, Frankfurt, 1978a; 5(3):142-150.

Dahlberg I. Ontical structures and universal classification. Bangalore: Sarada Ranganthan Endowment, 1978b. 64 p.

Dahlberg I. Towards estabilishment of compatibility between indexing languages. Internacional Classification. 1981; 8(2): 88-91.

Dahlberg I. Conceptual compatibility of ordering systems. Internacional Classification. 1983; 10(2):5-8.

Dávila AMR. Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, Mazzoni CJ, Wagner G, Pires PF, Grisard E C. GARSA: Genomic Analysis Resources for Sequence Annotation. Bioinformatics. 2005.

De Bruijn J, Ehrig M, Feier C. Ontology mediation, merging and aligning. In: John Davies, Paul Warren, and Rudi Studer: Semantic Web Technologies, John Wiley & Sons, 2006.

Dervin B. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In: Glazier J, Powell R (editors), Qualitative research in information management. Englewood, CO: Libraries Unlimited, 1992. p.61-84.

Doan A, Madhavan J, Domingos P, Halevy A. Learning to map between ontologies on the semantic web. Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, may, 2002, p. 662-673.

Euzenat J, Shvaiko P. Ontology matching. Springer Verlag, Berlin, Heidelberg (Germany), 2007.

Falbo RA. Integração de conhecimento em um ambiente de desenvolvimento de software. Rio de Janeiro: COPPE/UFRJ, 1998. (Tese apresentada à COPPE/UFRJ para obtenção do grau de Doutor em Ciências (D.Sc.) 81f. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1998.

Felicíssimo CH, Breitman KK. Taxonomic Ontology Alignment - an Implementation. Proceedings of the 7th International Workshop on Requirements Engineering, Tandil, 2004. p. 52-163.

Fernández M, Gómez-Pérez A, Juristo N. Methontology: from ontological art towards ontological engineering. Spring Symposium Series. Stanford. 1997. p. 33-40.

Gangemi A. Steve G, Giancomelli F. ONIONS: an ontological methodology for taxonomic knowledge integration. ECAI-96 Workshop on Ontological Engineering, Budapest, Aug. 13, 1996.

Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res. 2001; 11(8):1425-1433.

Glushkov VM, Skorokhod'ko EF, Strongnii AA. Evaluation of the degree of compatibility of information retrieval languages of document retrieval systems. Autom Doc & Math Ling. 1978;12(1):18-26.

GO (org.). Portal da Gene Ontology. Disponível em: <http://www.geneontology.org.>, acesso em: 24 abr. 2008.

Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. 1993; 5: 199-220.

Guarino N. Formal ontology and information systems. In: FOIS '98, 1, 1998, Trento, Italy. Proceedings Amsterdam: IOS Press; Tokyo: Omsha, 1998a. p. 3-15.

Guarino N, Carrara M, Giaretta P. An ontology of meta-level categories. LADSEB-CNR Int. Rep. 6/93, Preliminary version, nov. 1993.

HGP. Human Genome Program, U.S. Department of Energy, Genomics and its Impact on Science and Society: A 2003 Primer, 2003.

Hjørland B. Domain analysis in information science: eleven approaches – traditional as well as innovative. Journal of Documentation. 2002; 58(4): 422– 62.

Kalfoglou Y, Schorlemmer M. Ontology mapping: the state of the art. The Knowledge Engineering Review. 2003; 18(1): 1–31.

Lancaster FW. Vocabulary control for information retrieval. 2nd ed. Arlington, VA: Information Resources Press, 1986.

Latour B. Ciência em ação: como seguir cientistas e engenheiros sociedade afora. São Paulo: Editora Unesp, 1997.

Mcguinness D, Fikes R, Rice J, Wilder S. The Chimaera Ontology Environment. Proceedings of the The Seventeenth National Conference on Artificial Intelligence (AAAI 2000), Austin, Texas, Jul. 30-Aug. 3, 2000.

Mendes PN. Uma Abordagem para Construção e Uso no Suporte à Integração e Análise de Dados Genômicos. 2005. Dissertação (Mestrado em Programa em Pós-Graduação em Informática) - Núcleo de Computação Eletrônica - UFRJ, Rio de Janeiro, 2005.

Miller GA. WordNet: An on-line lexical database. Special issue of the International. Journal of Lexicography. 1990; 3(4).

Mougin F, Burgun A, Bodenreider O. Mapping data elements to terminological resources for integrating biomedical data sources. BMC Bioinformatics. 2006; 24(7) Suppl 3:S6.

Neville HH. Feasibility study of a scheme for reconciling thesauri covering a commom subject. Journal Doc. dec. 1970 ; 4(26) :313-36.

Neville HH. Thesaurus reconciliation. Aslib Proc. nov. 1972; 11(24): 620-6.

Noy NF, Musen MA. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, 2000, p. 450-455.

OBO. Open Biomedical Ontologies, 2005. Disponível em: <http://obo.sourceforge.net.>. Acesso em: 17 maio 2008.

OWL - Web Ontology Language, 2008. Disponível em: < http://www.w3.org/TR/owl-ref/>. Acesso em: 17 maio 2008.

Pinto S, Martins JP. A Methodology for Ontology Integration. Proceedings of First International Conference on Knowledge Capture, K-CAP 2001, Victoria, B.C., Canada, ACM Press, 2001.

Ranganathan SR. Prolegomena to Library Classification. New York: Asia Publishing House, 1967.

Reynaud C, Safar B. Exploiting WordNet as Background Knowledge. In: International ISWC'07 Ontology Matching (OM-07) Workshop, Busan, Corea, 2007.

Sabou M, D'aquin M, Motta E. Using the semantic web as background knowledge for ontology mapping. In: 1st International Workshop on Ontology Matching (OM-2006) at ISWC-2006, Athens, Georgia (USA), nov. 2006.

Sales LF. Ontologias de domínio: estudo das relações conceituais e sua aplicação. 141f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense, Rio de Janeiro, 2006.

Silva VS. Alinhamento de ontologias através do algoritmo de Alinhamento Local de Caminhos. Projeto Final de Graduação em Informática – Universidade Federal do Rio de Janeiro, Instituto de Matemática. 2008.

Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. In: Hájek Petr, Valdés-Villanueva Luis, Westerståhl Dag (eds.): Logic, Methodology and Philosophy of Science. Proceedings of the 12th International Conference, King´s College Publications, London, 2005, p. 505-520.

Soergel D. Compatibility of vocabularies. In: RIGGS, F.W. ed. The conta Conference; Proceedings of conference on conceptual and terminological anlysis in the social sciences. Bielefeld, may 24-7, 1981. Frankfurt, INDEKS Verl., 1982. p. 209-23.

Stumme G, Madche A. FCA-Merge: Bottom-up merging of ontologies. In: 7th Intl. Conf. on Artificial Intelligence (IJCAI '01), Seattle, WA, 2001, p. 225-230.

Su X. Semantic Enrichment for Ontology Mapping. PhD thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491, Trondheim, Norway, 2004.

Swartout W, Tate A. Guest editors' introduction: ontologies. IEEE Intelligent Systems. jan. 1999; 14(1): 18-9.

Vickery BC. Ontologies. J Info Sci, London. 1997; 23(4):227-86.

Wagner G. Geração e análise comparativa de seqüências genômicas de Trypanosoma rangeli. Dissertação (Mestrado em Biologia Celular e Molecular) - Fundação Oswaldo Cruz. 2006.

Weinstein PC. Ontology-Based Metadata: transforming the MARC Legacy. Digital Libraries, Pittsburg. 1998; p. 254-263.

Sobre os autores

Maria Luiza de Almeida Campos

Doutora em Ciência da Informação pelo Instituto Brasileiro em Informação Científica e Tecnológica - IBICT/UFRJ, com Pós-Doutorado no Laboratório de Biologia Molecular de Tripanosamatídeos e Flebotomídeos do Instituto Oswaldo Cruz – FIOCRUZ, pesquisando na área de ontologias genômicas, Professora Adjunta do Departamento de Ciência da Informação da Universidade Federal Fluminense e do Programa de Pós-Graduação em Ciência da Informação UFF. Possui atividades de ensino e pesquisa na área de Organização e Recuperação da Informação, Taxonomia; Ontologia, Construção de Tesauros. Atuou também como professora convidada de cursos de pós-graduação strictu sensu da Pós-Graduação em Informática da UFRJ (2002-2004) e latu-sensu em nível de aperfeiçoamento (Curso de Indexação, ano 1998-2000/USU; Curso de Gestão do Conhecimento, ano 1998/USU; Curso de Tesauro, ano 1994/UFF; Curso de Teoria da Classificação, ano1990/UNIRIO), e em nível de especialização (Curso em Planejamento, Organização e Direção de Arquivos - A Gestão da Informação, ano de 1996, 2007). Foi membro da Comissão Nacional de Princípios Terminológicos da Associação Brasileira de Normas Técnicas-ABNT. Desenvolve a pesquisa "Integração de Ontologias: O domínio da Bioinformática e a problemática da compatibilização terminológica, como bolsista em produtividade pelo CNpq. É coordenadora do grupo de pesquisa "Ontologia e Taxonomia, aspectos teóricos e metodológicos". Vêm atuando em diversas Instituições como consultora em atividades de elaboração de taxonomias, tesauros e de política de indexação, como Finep; Casa de Rui Barbosa; Fiocruz; SESC; Iphan; Central Globo de Produções e Petrobrás. É autora do livro "Linguagens Documentárias: teorias que fundamentam sua elaboração" e de artigos publicados em periódicos nacionais e internacionais.

Maria Luiza Machado Campos

Pesquisadora e professora do Departamento de Ciência da Computação do Instituto de Matemática da Universidade Federal do Rio de Janeiro. Formada em Engenharia Civil pela Universidade Federal do Rio Grande do Sul, é Mestre em Engenharia de Sistemas e Computação pela Coppe, Universidade Federal do Rio de Janeiro e PhD em Sistemas de Informação pela University of East Anglia, Norwich, Inglaterra. É bolsista pesquisadora nível 2 do CNPq. Suas áreas de atuação incluem bancos de dados, gestão do conhecimento, data warehousing, gerência de metadados e ontologias, aplicadas em especial aos domínios de bioinformática, petróleo e emergências. É membro ativo da comunidade de pesquisa em computação, tendo publicado diversos artigos em periódicos nacionais e internacionais e conferências na área, orientado dezenas de dissertações, teses de mestrado e doutorado, e coordenado projetos de pesquisa financiados pela Finep, CNPq e Faperj. Tem também atuado em projetos de consultoria junto a empresas, na implantação de tecnologias de ponta voltadas para a gerência, integração e exploração de informações nas organizações.