

Integração de Ontologias: o domínio da Bioinformática



Maria Luiza de Almeida Campos

Departamento de Ciência da Informação – Universidade Federal Fluminense.
Programa de Pós-Graduação em Ciência da Informação UFF-IBICT, Niterói, Brasil
marialuizalmeida@gmail.com

Resumo

O crescimento da utilização de arquiteturas distribuídas, especialmente no ambiente da Web, contribuiu para que informações originalmente isoladas sejam disponibilizadas para serem utilizadas de maneira integrada. Ontologias assumem papel fundamental nessa integração, viabilizando a interoperabilidade semântica de sistemas distribuídos heterogêneos. Este artigo é fruto da pesquisa “Integração de Ontologias: o domínio da Bioinformática e a problemática da compatibilização terminológica”, que tem por objetivo apresentar diretrizes que permitam o desenvolvimento, uso e integração de ontologias empregadas na descrição e recuperação dos recursos e serviços de Bioinformática.

Palavras-chave

Integração de ontologia, ontologia de domínio, bioinformática, base teórica e metodológica em ontologia, compatibilização e integração de linguagem

Este artigo tem por objetivo apresentar a pesquisa em desenvolvimento desta autora, apoiada pelo CNPq (período 2005/2008), onde propõe investigar mecanismos de uso, desenvolvimento e integração de ontologias no domínio da Bioinformática, especificamente no campo que envolve as pesquisas em Genoma e Transcriptoma, visando apoiar os estudos nessa área, que estão sendo desenvolvidos por instituições como a Fiocruz, trazendo aspectos teóricos e metodológicos da Ciência da Informação no seu domínio de competência relacionado à elaboração de linguagens documentárias e organização do conhecimento.

Especificamente, pretendemos identificar nas bases teóricas da Ciência da Informação e da Terminologia,

no campo específico de construção e compatibilização de linguagens, propostas teóricas e metodológicas que possibilitem o desenvolvimento, o uso e a integração de ontologias.

No que tange à integração de ontologias, na primeira etapa desta pesquisa, pretendemos identificar as Ontologias do domínio de Genoma e Transcriptoma existentes internacionalmente e nacionalmente através de ações que permitam a comparação dos diversos modelos taxonômicos empregados para a representação dos domínios e de suas relações, identificando os níveis de compatibilização semântica e lingüística, visando a apresentação de proposta de harmonização terminológica, evidenciando diretrizes para a integração de ontologias.

No que concerne aos princípios teóricos e metodológicos que venham a auxiliar a elaboração de ontologias, na segunda etapa desta pesquisa, pretendemos identificar no Consórcio Internacional que aspectos do projeto não estão sendo contemplados nas diversas ontologias analisadas, através da aplicação do método de compatibilização de linguagens, visando atender às especificidades brasileiras. Os domínios temáticos não atendidos serão então definidos como espaço empírico para o desenvolvimento de propostas metodológicas para a sua modelização através da identificação de bases teóricas e metodológicas estudadas e sistematizadas para elaboração de ontologias.

No âmbito dos Sistemas de Recuperação de Informação, a organização e a recuperação de informações sempre estiveram condicionadas à tecnologia associada. Atualmente, bases de dados de todos os tipos têm proliferado com a disponibilização de informações em rede e, principalmente, na Web. A recuperação dos conteúdos informativos ainda não é realizada de forma satisfatória, devido à falta de ferramentas de acesso adequadas, que viabilizem, por exemplo, o controle terminológico.

Para garantir esta precisão, verifica-se a necessidade de ferramentas taxonômicas e terminológicas para o tratamento semântico de informações contidas em bases de dados, viabilizando, entre outros processos, a integração de informações como auxílio ao desenvolvimento de pesquisa em domínios de conhecimento.

Ferramentas semânticas como ontologias precisam ser construídas em meio informatizado, abarcando domínios de estudos e pesquisa em língua portuguesa, para serem utilizadas em bibliotecas digitais e virtuais, em sistemas para gestão de conhecimento, para viabilizarem processos de integração de informações entre pesquisadores, como auxílio para as ferramentas de busca de um modo geral, e, principalmente, como um instrumento para a melhoria do tratamento e da recuperação de informação na rede.

Ontologia é um conjunto de conceitos padronizados onde termos e definições devem ser aceitos por uma comunidade no âmbito de um domínio, e tem por finalidade permitir que múltiplos agentes compartilhem conhecimento. Uma ontologia consiste em termos, definições e axiomas relativos a eles. As ontologias constituem um meio poderoso de inter-relacionar sistemas. São elaboradas, principalmente, visando à estruturação de bases de conhecimento ou para serem utilizadas como ferramentas semânticas no suporte à interoperabilidade entre sistemas de informação.

Atualmente, o crescimento da utilização de arquiteturas distribuídas, especialmente no ambiente da Web, de interfaces abertas de acesso a bancos de dados, de tecnologias de mediadores e de padrões de formato para troca de dados contribuiu para que informações originalmente isoladas sejam disponibilizadas para serem utilizadas de maneira integrada. Ontologias assumem papel fundamental nesta integração, viabilizando a interoperabilidade semântica de sistemas distribuídos heterogêneos. Desta forma, ontologias estabelecem fun-

damentos de significados conceituais sem os quais a Web Semântica não seria possível, devido à heterogeneidade dos conceitos representados (JACOB, 2003).

A heterogeneidade tem sido identificada como um dos problemas mais importantes e difíceis de tratar. Ela envolve a interoperabilidade e cooperação entre múltiplas fontes de informação, retratando diferenças sintáticas, semânticas e estruturais entre sistemas.

A heterogeneidade semântica representa atualmente o maior empecilho para a interoperabilidade semântica, representando um grande desafio para a integração de informações na Web. Para tratar esse problema, é preciso buscar uma linguagem capaz de representar conhecimento e regras, além de inferir novos dados.

Isso se dará a partir de inter-relacionamentos entre ontologias específicas de domínios, que têm como premissas o uso racional de metadados, para descrição de dados de forma homogênea, o uso sistemático de ontologias, preenchendo a lacuna entre fontes de dados heterogêneas, e a utilização de associações semânticas, tratando a interoperabilidade entre domínios (ADAMS, 2002).

Dentro do domínio de desenvolvimento de ontologias, as abordagens para a sua construção são específicas e limitadas. A literatura, no âmbito da Ciência da Computação, tem privilegiado ora as ontologias como vocabulários de domínios específicos, sem um suporte teórico, ora um conjunto de regras e aportes teóricos, sem elementos que orientem a construção de vocabulários que permitam a elaboração de definições lógicas. Além disto, verifica-se a existência de dois grandes problemas nas metodologias (FERNÁNDEZ-LOPEZ, 1999; GUARINO, 2002; JONES, 1998; SURE, 2002) para projeto de ontologia: a falta de explicação sistemática de como e onde serão usadas as abordagens teóricas dentro de seu processo de elaboração; a não existência dos estágios de integração e manutenção de ontologia no método na maioria das metodologias.

Acredita-se que o aporte teórico e metodológico existente no âmbito da Ciência da Informação se beneficiando de estudos no escopo da Teoria da Terminologia (WUESTER, 1981), Teoria do Conceito (DAHLBERG, 1978), Teoria da Classificação (RANGANATHAN, 1967) e Compatibilização de Linguagens (NEVILLE, 1970; DAHLBERG, 1981, 1983) possa apresentar propostas eficazes de aplicação. Por outro lado, estas áreas podem se beneficiar atuando numa área bastante aplicada da questão, fugindo da complexidade de um tratamento excessivamente formal. Espera-se que este projeto venha possibilitar maior integração das informações, criando mecanismos para consolidar a participação do Brasil nos consórcios internacionais que envolvem a pesquisa em Genoma e Transcriptoma.

As pesquisas em Bioinformática no Brasil vêm desenvolvendo estudos que têm por finalidade prover um ambiente que possa oferecer informação semântica sobre os recursos científicos, como dados e programas nessa área, e possibilitar o uso desses recursos de forma conjunta pela comunidade científica interessada. Um

dos aspectos que envolvem a descrição e a recuperação desses recursos e serviços está relacionado ao desenvolvimento de uma linguagem padronizada e consensual para facilitar o entendimento dos vocabulários, muitas vezes interdisciplinares.

No campo da genômica, iniciativas da comunidade científica internacional, nos últimos anos, levaram a um crescimento explosivo de informações biológicas geradas todos os dias (HGP, 2003). A preocupação inicial era 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 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., visando responder a perguntas, tais como: Qual é a proteína que este gene codifica? Qual a função desta proteína neste organismo? Este gene é similar a outro gene presente em organismo distinto? (MENDES, 2005).

Desta forma, é importante considerar a relevância da gerência, descrição e organização dos recursos científicos em meio digital para a pesquisa em Bioinformática. Especificamente, nesta área, nem sempre esses recursos estão disponíveis para o biólogo, e muitas vezes este tem que recorrer à utilização de programas proprietários, residentes em outras instituições. Assim, para que essas equipes e/ou instituições troquem recursos científicos entre si, é preciso encontrar uma forma comum de descrição e acesso a esses recursos, de modo a facilitar a busca e a integração dos mesmos.

Assim, 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 a possibilitar com consistência a recuperação de informações. Iniciativas no campo do tratamento terminológico têm sido apresentadas através de repositórios de ontologias.

A Biblioteca de Ontologias OBO – *Open Biological Ontologies* é um repositório de terminologias desenvolvido para uso compartilhado entre diversos domínios biológicos e médicos. Dentre os mais difundidos vocabulários componentes da OBO, podemos destacar a *Gene Ontology* (GO). A GO compreende termos referentes a três categorias de assunto, ou seja, componentes celulares, processos biológicos e funções celulares.

No Brasil, especificamente na 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”, atualmente

financiada pelo CNPq e coordenado pelo Dr. Alberto M. R. Dávila da 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 desses recursos de forma conjunta pela comunidade científica interessada, e vem utilizando a GO para as anotações em seu banco de dados. No nível internacional, como apresentado anteriormente, reconhece-se a Gene Ontology, que inclui termos referentes a processos biológicos, componentes celulares e funções moleculares, de maneira independente de espécies, entre outras iniciativas. Entretanto, até o momento não se identificam, em níveis nacional e internacional, ontologias desenvolvidas dentro do recorte conceitual específico, ou seja, de Genoma e Transcriptoma para atender às demandas dos grupos coordenados pela Fiocruz. Apesar dos esforços internacionais, a *Gene Ontology* parece não possuir classes de conceitos que venham a atender plenamente as pesquisas desenvolvidas no Brasil; em alguns casos é necessário investigar a harmonização existente entre termos e o seu conteúdo conceitual.

Como resultados esperados do projeto, podemos citar: diretrizes para integração de ontologias, envolvendo questões relativas às definições conceituais; diretrizes para o desenvolvimento de ontologias; modelização do domínio de Genoma e Transcriptoma. Neste último aspecto, estudos relacionados à modelização de domínios, princípio fundamental para a etapa de elaboração de taxonomias para ontologias, já vem sendo ponto de investigação em nossos estudos e pesquisa (CAMPOS, 2004).

O projeto envolve pesquisadores, professores e alunos de programas de pós-graduação e cursos de graduação de instituições de renome nas áreas de conhecimento associadas ao tema do projeto, garantindo com isso a formação de recursos humanos capacitados em temáticas estratégicas para o tratamento, integração e recuperação de informações, mas ainda de pouca divulgação no país.

Nesta primeira etapa do projeto pretendemos atingir aos seguintes objetivos: 1. revisão de literatura no domínio da Ciência da Informação, Ontologia e Terminologia; 2. análise e identificação de princípios para a integração e desenvolvimento de ontologias; 3. levantamento e análise de Ontologias no domínio de Genoma e Transcriptoma.

Atualmente, estamos nos concentrando nas seguintes atividades: 1. revisão da literatura sobre integração e compatibilização de linguagens no âmbito da Ciência da Informação, Ciência da Computação e Terminologia. Por meio do levantamento realizado, estamos elaborando um banco de dados, onde as informações estão sendo tratadas, possibilitando acesso ao grupo de pesquisadores envolvidos; 2. levantamento de ontologias existentes no domínio de Genoma e Transcriptoma, com a finalidade de mapear as áreas e subáreas dentro dos domínios apresentados e auxiliar os pesquisadores na identificação da produção nestes domínios; 3. estudos referentes às definições conceituais.

No caso das ontologias, as definições propiciam a possibilidade de compatibilização semântica, pois descrevem o conteúdo semântico do signo lingüístico (o termo). Esta descrição possibilita que agentes inteligentes possam entender o significado de um termo e estabelecer inferências sobre esses significados, pois a definição é composta de características de conceitos, que são também conceitos que se relacionam formando o entendimento semântico dos termos em questão.

Desta forma, as definições são de fundamental importância para a elaboração de ontologias consistentes. Entretanto, é fato que as ontologias existentes se ressentem hoje de um padrão definitório para a sua elaboração. Isto é bastante problemático quando se coloca a questão da compatibilização de linguagens que operam em bases cooperativas, como é o caso das pesquisas que vêm sendo desenvolvidas em Bioinformática. Este estudo tem por pretensão teórica e prática aproximar campos de atividades diferentes em torno da problemática definitória e da compatibilização de linguagens.

Até a década de 1960 as definições tinham um caráter mais conceitual, filosófico, evidenciavam o que pensar acerca de um conceito. Entretanto, para atender às necessidades da Ciência, que possui um caráter menos filosófico, mais operacional, foi necessário um esforço teórico que visasse a elaboração de definições operacionais, que pretendiam relacionar um dado conceito, além de conceitos que indicavam o gênero próximo e a diferença específica, mas também, a outros conceitos que procurava definir certas operações onde o conceito seria aplicado, ou melhor, ao que seria observado se determinadas operações fossem executadas.

Por outro lado, a questão das definições é também um campo de estudo no âmbito da Terminologia desde a década de 1930, com (WUESTER, 1981), que tinha por objetivo o estudo do termo no âmbito da língua de especialidade, ou seja, da própria ciência.

Segundo DAHLBERG (1983a) o conceito de definição pode ser apresentado como: “a equivalência entre um *definiendum* (o que deve ser definido) e um *definiens* (como algo deve ser definido) com o propósito de delimitar o entendimento do *definiendum* em qualquer caso de comunicação.”

A partir desta explicação, apresenta três tipos de definições: nominal, ostensiva e conceitual. A definição nominal é aquela onde o *definiendum* é uma expressão verbal e o *definiens* é uma equivalência textual deste termo, como, por exemplo, A = B. A definição ostensiva é aquela onde o *definiens* é estabelecido apontando-se para o referente nomeado pelo *definiendum*, ou seja, C = A. Já a definição conceitual, também denominada definição real, ocorre quando o *definiens* contém as características necessárias de um referente nomeado pelo *definiendum*, ou seja, C = B de A.

No caso das questões que envolvem a compatibilização e a integração entre ontologias, nos interessa o estudo das definições conceituais e das nominais, pois estas permitem a compatibilização no plano semântico e no plano lingüístico, respectivamente.

A partir do apresentado, como contribuição importante do projeto, defendemos ainda a convergência de métodos e técnicas de duas áreas de conhecimento fundamentais ao desenvolvimento de práticas relativas às ontologias: Ciência da Informação e Ciência da Computação. Além de um espaço empírico de aplicação, ou seja, o domínio da Bioinformática. Na maior parte dos projetos neste tema, pode-se observar um viés específico de uma dessas áreas, sem considerar importantes contribuições que a outra área poderia trazer. Das interações anteriores dos pesquisadores, resultou a firme convicção da importância de uma abordagem integrada e multidisciplinar no tratamento do tema ontologia.

Referências bibliográficas

ADAMS, K. The semantic web: differentiating between taxonomies and ontologies. *On line*, v.26, n.4, p.20-23, Jul.-Aug., 2002.

CAMPOS, M. L. A. Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. *Ciência da Informação*, Brasília, v.33, n.1, 2004.

DAHLBERG, I. A referent-oriented analytical concept theory of interconcept. *International Classification*, v.5, n.3, p.142-150, 1978.

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

DAHLBERG, I. Terminological definitions: characteristics and demands. In: *Problèmes de la définition et de la synonymie en terminologie*. Québec, GIRSTERM, 1983a. p. 15-34.

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

FERNÁNDEZ-LÓPEZ, M. Overview of methodologies for building ontologies. In: *IJCAI-99 Workshop on ontologies and problem-solving methods (KRR5)*, 1999, Stockholm, *Procedins...* Stockholm, Sweden, 1999.

HGP. Human Genome Program, US. Department of Energy, Genomics and its impact on science and society: A 2003 Primer, 2003. Disponível em: <http://www.ornl.gov/sci/techresources/HumanGenome/publicat/primer2001/index.sht>

HAMMOND, W. Dimensions in compatibility. In: NEWMAN, S. M. (Ed.). *Information systems compatibility*. Washington: Spartan Books, 1965. p. 7-17.

JACOB, E.K. Ontologies and the semantic web. *Bulletin of the American Society for Information Science and Technology*, Apr./may. 2003.


JONES, D.; BENCH-CAPON, T.; VISSER, P. Methodologies for ontology development. In: *IT&Knows conference of the 15th IFIP world computer congress*, 1998, Budapest, *Proceedings...* Budapest, Bulgaria: Chapman-Hall, 1998.

MENDES, P.N. **Uma abordagem para a construção e uso de ontologias no suporte à integração e análise de dados genômicos**. Dissertação (Mestrado em Ciência da Computação), Instituto de Matemática/NCE da UFRJ. Rio de Janeiro: Brasil, 2005.

NEVILLE, H. H. Feasibility study of a scheme for reconciling thesauri covering a common subject. **Journal of Documentation**, n.26, v.4, p.313-36, dec. 1970.

RANGANATHAN, S. R. **Prolegomena to library classification**. Bombay: Asia Publishing House, 1967.

SURE, Y.; STAAB, S.; STUDER, R. Methodology for development and employment of ontology based knowledge management applications. **SIGMOD Record**, v.31, n.4, p.18-23, 2002

WUESTER, E. L'étude scientifique générale de la terminologie, zone frontalière entre la Linguistique, la Logique, l'Ontologie, l'Informatique et les Sciences des Choses. In: RONDEAU, G.; FELBER, F. (Org.). **Textes choisis de terminologie: I: fondements théoriques de la terminologie**. Québec: GIRSTERM, 1981. p.57-114. 

Sobre a autora

Maria Luiza de Almeida Campos

Doutora em Ciência da Informação pelo Instituto Brasileiro em Informação Científica e Tecnológica - IBICT/UFRJ, Professora Adjunta e Chefe do Departamento de Ciência da Informação da Universidade Federal Fluminense e do Programa de Pós-Graduação em Ciência da Informação UFF/IBICT. Possui atividades de ensino e pesquisa na área de Organização e Recuperação da Informação, Taxonomia; Ontologia, Construção de Tesouros. 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 Tesouro, ano 1994/UFF; Curso de Teoria da Classificação, ano 1990/UNIRIO), e em nível de especialização (Curso em Planejamento, Organização e Direção de Arquivos - A Gestão da Informação, ano de 1996, 2007). Foi membro, desde a sua criação em 1992, da Comissão Nacional de Princípios Terminológicos da Associação Brasileira de Normas Técnicas-ABNT, que tem por objetivo a elaboração de normas terminológicas nacionais. Desenvolve a pesquisa "Integração de Ontologias: o domínio da Bioinformática e a problemática da compatibilização terminológica, como bolsista em produtividade pelo CNPq. Além disso, é coordenadora do grupo de pesquisa registrado pelo CNPq "Ontologia e Taxonomia, aspectos teóricos e metodológicos". Atualmente atua em diversas Instituições como consultora em atividades de elaboração de taxonomias, tesouros 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.